1465

NAVAL RESEARCH LOGISTICS QUARTERLY

DEPOSITORY
20 AUG 1980

SEP 4 1980

SEPTEMBER 1980
VOL. 27, NO. 3

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

407-B

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORIAL BOARD

Marvin Denicoff, *Office of Naval Research*, Chairman

Murray A. Geisler, *Logistics Management Institute*

W. H. Marlow, *The George Washington University*

Ex Officio Members

Thomas C. Varley, *Office of Naval Research* Program Director

Seymour M. Selig, *Office of Naval Research* Managing Editor

MANAGING EDITOR

Seymour M. Selig
Office of Naval Research
Arlington, Virginia 22217

ASSOCIATE EDITORS

Frank M. Bass, *Purdue University*
Jack Borsting, *Naval Postgraduate School*
Leon Cooper, *Southern Methodist University*
Eric Denardo, *Yale University*
Marco Fiorello, *Logistics Management Institute*
Saul I. Gass, *University of Maryland*
Neal D. Glassman, *Office of Naval Research*
Paul Gray, *Southern Methodist University*
Carl M. Harris, *Center for Management and Policy Research*
Arnoldo Hax, *Massachusetts Institute of Technology*
Alan J. Hoffman, *IBM Corporation*
Uday S. Karmarkar, *University of Chicago*
Paul R. Kleindorfer, *University of Pennsylvania*
Darwin Klingman, *University of Texas, Austin*
Kenneth O. Kortanek, *Carnegie-Mellon University*
Charles Kriebel, *Carnegie-Mellon University*
Jack Laderman, *Bronx, New York*
Gerald J. Lieberman, *Stanford University*
Clifford Marshall, *Polytechnic Institute of New York*
John A. Muckstadt, *Cornell University*
William P. Pierskalla, *University of Pennsylvania*
Thomas L. Saaty, *University of Pittsburgh*
Henry Solomon, *The George Washington University*
Wlodzimierz Szwarc, *University of Wisconsin, Milwaule*
James G. Taylor, *Naval Postgraduate School*
Harvey M. Wagner, *The University of North Carolina*
John W. Wingate, *Naval Surface Weapons Center, Whit Oa*
Shelemyahu Zacks, *Virginia Polytechnic Institute and State University*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistic an will publish research and expository papers, including those in certain areas of mathematics, statistics, and econcics relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, ne September, and December and can be purchased from the Superintendent of Documents, U.S. Government Pr.ting Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cc t of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the Cfice of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regula ns, P-35 (Revised 1-74).

ON THE RELIABILITY, AVAILABILITY AND BAYES CONFIDENCE INTERVALS FOR MULTICOMPONENT SYSTEMS

William E. Thompson

Columbia Research Corporation
Arlington, Virginia

Robert D. Haynes

ARINC Research Corporation
Annapolis, Maryland

ABSTRACT

The problem of computing reliability and availability and their associated confidence limits for multi-component systems has appeared often in the literature. This problem arises where some or all of the component reliabilities and availabilities are statistical estimates (random variables) from test and other data. The problem of computing confidence limits has generally been considered difficult and treated only on a case-by-case basis. This paper deals with Bayes confidence limits on reliability and availability for a more general class of systems than previously considered including, as special cases, series-parallel and standby systems applications. The posterior distributions obtained are exact in theory and their numerical evaluation is limited only by computing resources, data representation and round-off in calculations. This paper collects and generalizes previous results of the authors and others.

The methods presented in this paper apply both to reliability and availability analysis. The conceptual development requires only that system reliability or availability be probabilities defined in terms acceptable for a particular application. The emphasis is on Bayes Analysis and the determination of the posterior distribution functions. Having these, the calculation of point estimates and confidence limits is routine.

This paper includes several examples of estimating system reliability and confidence limits based on observed component test data. Also included is an example of the numerical procedure for computing Bayes confidence limits for the reliability of a system consisting of N failure independent components connected in series. Both an exact and a new approximate numerical procedure for computing point and interval estimates of reliability are presented. A comparison is made of the results obtained from the two procedures. It is shown that the approximation is entirely sufficient for most reliability engineering analysis.

INTRODUCTION

The problem of computing reliability, availability, and confidence limits for multicomponent systems where some or all of the component reliabilities and availabilities are statistical estimates from test and other data has appeared often in the literature. The problem of computing these confidence limits has generally been considered difficult and treated only on a case by case basis. The present paper deals with Bayes confidence limits on reliability and steady state availability for a general class of fixed mission time, two-state systems including, as special cases, series-parallel, stand-by and others that appear in the applications. Further, a fixed mission length is assumed. It is also assumed that neither reliability growth nor deterioration occur during the life of the system and the system becomes as good as new after each repair. Finally, we assume that no environmental changes, which could affect reliability occur. The posterior

distributions obtained are exact in theory and their numerical evaluation is limited only by computing resources, data representation and round-off in calculation. The present paper collects and generalizes previous results of the authors and others.

The methods obtained in the following apply both to reliability and steady state availability analysis and to avoid repeated reference to "reliability or availability", the discussion references only reliability with the understanding that the terms system reliability R and component reliability r_i can be replaced by system availability A and component availability a_i. The conceptual development requires only that R and A be probabilities defined in terms acceptable for a particular application. The emphasis is on the determination of the posterior distribution functions. Having these, the calculation of point estimates and confidence intervals is routine.

BAYES CONFIDENCE INTERVALS

In the Bayes inference model, the unknown probability, R, $0 \leqslant R \leqslant 1$, is considered a random variable whose posterior density is the result of combining prior information with test data to obtain a probability density function $f(R)$ for R. If the posterior density of R is seen to be spread out, then relatively more uncertainty in the value of R obtains than when the posterior density is concentrated closely about some particular value. The posterior density function provides the most complete form of information about R, but sometimes summary information is desired. A point estimate is one such form of summary information and this can be selected in various ways and is analogous to the familiar statistical problem of characterizing an entire population by some parameter value. Examples are mean, mode, median, etc. A point estimate has the disadvantage of ignoring the information concerning the uncertainty in the unknown reliability. Confidence intervals derived from $f(R)$ provide such additional information. The true but unknown (and unknowable except with infinite data) reliability R_0 is some specific value of the random variable R, $0 \leqslant R \leqslant 1$. Conceptually, R_0 can be considered a random sample from $0 \leqslant R \leqslant 1$ made when the system was built. We can never know that R_0 is, but $f(R)$ gives a measure of the likelihood that $R_0 = R$ for each $0 \leqslant R \leqslant 1$. If $F(R) = \int_0^R f(R)\,dR$ denotes the distribution function of R then

$$\text{Prob}\,\{R_1 \leqslant R_0 \leqslant R_2\} = F(R_2) - F(R_1)$$

and $[R_1, R_2]$ is an interval estimate of R of confidence $c = F(R_2) - F(R_1)$. The interpretation is simply that, based on the prior and current data the probability is c that the unknown system reliability lies between R_1 and R_2. The interval $[R_1, R_2]$ has been called [25] a Bayes c level confidence interval. For $R_2 = 1$, R_1 is called the lower c level confidence limit. For $R_1 = 0$, R_2 is called the upper c level confidence limit. Given $f(R)$ and $F(R)$, Bayes confidence limits for any c can be obtained by graphical or numerical methods and the procedure is generally not difficult. Numerical examples and discussion of numerical methods are given in [25,27,8,26,28,29].

DEFINITION OF STRUCTURE FUNCTION

To establish the relationship between the reliabilities of the components of a system and the reliability of the entire system, the way in which performance and failure of the components affects performance and failure of the system must be specified. For this purpose, as in [5,10,15], the state of any component is coded 1 when it performs and 0 when it fails. The state of all N components of the system can then be coded by a vector of N coordinates

$$x = (x_1, x_2, \ldots, x_N)$$

where $x_i = 0$ means the i-th component fails and $x_i = 1$ means that it does not fail. All possible states of the system are represented by the 2^N different values this vector can assume.

Where an explicit mission time dependence is required, a random process $y(t) = \{y_1(t), \ldots, y_N(t)\}$ can be defined as in [15] so that to each component trajectory a measure x_i is assigned. Then, for example: $x_i = 1$ if $y_i(t)$ is a failure-free process over some interval $0 \leqslant T_{i1} \leqslant t \leqslant T_{i2}$, and $x_i = 0$ if at least one failure occurs.

Some of the 2^N states cause the system to fail and the others cause the system to perform. The response of the system as a whole is written as a function $\phi(x)$ of x such that $\phi(x) = 0$ when the system is failed in state x, and $\phi(x) = 1$ when the system performs in state x. This function $\phi(x)$ is known in the literature [5,10,23] and has been called a structure function of order N.

The structure function can be written in a systematic way for any series parallel system. When the system is not too large the structure function can also be written by observation for many more general systems. The structure function can always be written for a system of N components by enumeration of its 2^N states. For large systems this is at best very tedious, but generally short cuts can be found which simplify the process. The structure function is convenient for conceptual development of the theory and provides a very general notation which is why it is used here. What is required in the application of the present results is the formula for system reliability in terms of component reliabilities as is done in [25,27, and 8]. The structure function provides this formula in a general form but other methods are available. Some of these methods are identified and referenced in [17] along with a new and useful algorithm based on graph theory.

DEFINITION OF RELIABILITY FUNCTION

Assume that the components of the system are failure independent so that the elements of the state vector $x = (x_1, \ldots, x_N)$ are independent random variables with probability distributions

$$Pr\{x_i = 1\} = r_i$$
$$Pr\{x_i = 0\} = 1 - r_i$$

where r_i is the reliability of the i-th component.

The structure function $\phi(x)$ is also a random variable with

$$Pr\{\phi(x) = 1\} = R$$
$$Pr\{\phi(x) = 0\} = 1 - R$$

where R is the reliability of the system. R is the expected value of $\phi(x)$ so that

$$R = E\{\phi(x)\} = \sum \phi(x) r_1^{x_1} (1 - r_1)^{1-x_1} \ldots r_N^{x_N} (1 - r_N)^{1-x_N} \tag{1}$$

where the summation is over all 2^N states of the system.

In a particular application given the structure function and the values of all component reliabilities, the system reliability, R, can be computed explicitly using (1). References [5], [10], and [23] provide further discussion with examples of ϕ and R.

RELIABILITY ESTIMATION FROM TEST DATA

In many applications the system structure is known but some or all of the component reliabilities are unknown and must be estimated from tests and other data. As a result, statements concerning these component and system reliabilities are subject to the uncertainties of statistical estimation. A method of treating this uncertainty is provided by a Bayes analysis which considers the unknown component reliabilities as random variables and leads to Bayes confidence intervals for both component and system reliabilities. The following is an extension and generalization of previous analysis of this kind [25,27,8,7,14,23,29].

BAYES MODEL

Assume a system of N failure independent components has a known structure function $\phi(x)$ and reliability function $R(r)$ $r = (r_1, \ldots, r_N)$ of the form (1). Suppose that among the N separate components of the system some are known to have identical reliabilities say i,j, and k, for example, then since $r_i = r_j = r_k$, the symbols r_j and r_k can be replaced by r_i everywhere in (1). Finally, in this way there remain only $N' < N$ different r's, one of each reliability value. In addition, suppose that among the N' different component reliabilities $N'-n$ are known constants and thus there remain n different types of components with unknown reliabilities. By a simple change in notation these n different, unknown reliabilities are denoted by $p = (p_1, p_2, \ldots, p_n)$.

By multiplying out factors $(1 - p)$ and collecting terms, the system reliability (1) can then be written in the equivalent form

$$R(p) = \sum_j a_{o_j} p_1^{a_{1j}} \ldots p_n^{a_{nj}} \tag{2}$$

where the constants a_{ij}, are integer for $i \neq 0$.

Using a Bayes inference model, the unknown p_i are considered independent random variables with known posterior density functions,

$$f_i(p_i),\ 0 \leqslant p_i \leqslant 1,\ i = 1, \ldots, n.$$

The system reliability, $R(p)$ is then also a random variable, defined by (1) with unknown distribution function $H(R)$.

In applications, what is required is the calculation of $H(R)$ given the $f_i(p_i)$; $i = 1, 2, \ldots, n$. Having obtained $H(R)$, point estimates and confidence intervals on R can be obtained directly. This result is also required for risk, cost and other analyses based on the Bayes model. The method for an explicit numerical evaluation is presented in the following section.

EVALUATION OF THE POSTERIOR DISTRIBUTION

The proposed method of evaluating the posterior distribution function $H(R)$ is based on an expansion of $H(R)$ in Chebyshev polynomials of the second kind [1,16]. The main advantages of this method lie in the rapid convergence properties of the Chebyshev expansion and the convenient numerical computation for its evaluation. Although a description of the procedure has been presented in [8] and [7], for the sake of completeness, we shall outline the main steps below.

Expansion by Chebyshev Polynomials

Let $H(R)$ denote the posterior distribution,

$$H(R) = \int_0^R h(R)dR, \ 0 \leqslant R \leqslant 1,$$

where $h(R)$ is the posterior density of the reliability of the overall system. By definition, $H(R)$ satisfies the boundary conditions:

(3) $$H(0) = 0; \ H(1) = 1$$

Let us introduce a new function $Q(R)$ defined by

(4) $$Q(R) = H(R) - R$$

the $Q(R)$ satisfies the boundary conditions

(5) $$Q(0) = Q(1) = 0$$

and can be expanded in a Fourier sine series of the following form:

(6) $$Q(R) = \frac{4}{\pi} \sin\theta [b_0 + b_1 \frac{\sin 2\theta}{\sin\theta} + \ldots + b_k \frac{\sin (k+1)\theta}{\sin\theta} + \ldots]$$

where the angular variable θ is related to R by the relation

(7) $$R = \cos^2 \frac{\theta}{2}.$$

The coefficients b_k of the expansion (6) can be determined by:

(8) $$b_k = \int_0^1 [H(R) - R] \ U_k^* (R)dR$$

where $U_k^* (R) \equiv \frac{\sin (k+1)\theta}{\sin\theta}$ is the shifted Chebyshev polynomial of the second kind [1,16] which can be computed by the recursion relations:

(9) $$U_{k+1}^* (R) = (4R - 2) \ U_k^* (R) - U_{k-1}^* (R)$$

with

$$U_0^* (R) = 1 \quad U_1^* (R) = -2 + 4R$$
$$U_2^* (R) = 3 - 16R + 16R^2$$

If we express $U_k^* (R)$ explicitly as a kth order polynomial

(10) $$U_k^* (R) = \sum_{i=0}^{k} C_{ik} R^k$$

then Equation (8) becomes

(11) $$b_k = \sum_{i=0}^{k} C_{ik} \left\{ \int_0^1 R^i H(R)dR - \int_0^1 R^{i+1} dR \right\}.$$

It can be shown, integrating by parts, that

(12) $$M_i\{H(R)\} = \frac{1}{i+1} \{1 - M_{i+1}[h(R)]\}.$$

Thus, Equation (11) becomes

$$b_k = \sum_{i=0}^{k} C_{ik} \left\{ \frac{1-M_{i+1}[h(R)]}{i+1} - \frac{1}{i+2} \right\}. \tag{13}$$

Note that the Chebyshev coefficients C_{ik} can be computed independently of the moments. They may be stored in the form of a triangular matrix if sufficient storage space is available. A simple algorithm for recursively calculating the coefficients is $C_{i,k+1} = 4C_{i-1,k} - 2C_{i,k} - C_{i,k-1}$.

Computations and Results

To complete the analysis it remains to compute the moments of $h(R)$ given the density functions $f_i(p_i)$ and then use (13) to compute the b_r.

From (2) $R^k(p)$ $k = 1, 2, \ldots$ can be written as a finite sum

$$R^k(p) = \sum_j a_{ojk} p_1^{\,a_{1jk}} \ldots p_R^{\,a_{njk}}, \tag{14}$$

where the a_{ijk} are independent of the p_i and also integers for $i \neq 0$. Using this result and the fact that the expected value of a sum is the sum of the expected values and the expected value of a product of independent random variables is the product of the expected values, it follows that

$$M_k\{h\} = \sum_j a_{ojk} M_{a_{1jk}} \ldots M_{a_{njk}} \tag{15}$$

where $M_{a_{ijk}}$ denotes the a_{ijk} 'th moment of p_i.

Having determined the coefficients b_k we can write down the final expression for $H(R)$ from Equations (4) and (6) as follows:

$$H(R) = R + \frac{8}{\pi} \sqrt{R(1-R)} \{b_0 + b_1 U_1^*(R) + \ldots + b_k U_k^*(R) + \ldots\}. \tag{16}$$

This result is exact in the sense that the error can be made arbitrarily small by taking a sufficient number of terms. References [8] and [7] give a discussion of numerical considerations and examples. Generally, (16) has been found very convenient for numerical calculation using an electronic digital computer.

MODELS FOR APPLICATION

To evaluate $H(R)$ the posterior distribution $f_i(p_i)$ for each different component reliability p_i is required. The derivation of these require application of Bayes inference procedures on a case by case basis. The theory can be found in [20,4,19,2,3,24,6,18] and some specific applications in [25,27,8,7,14,1,16,12,23]. A tabulation for some familiar models of mathematical reliability theory is presented in the following.

Component With Constant Failure Rate

A single component has an unknown constant failure rate λ and fixed mission time t. Component reliability $p = \exp(-\lambda t)$ is regarded as a random variable. The natural conjugate prior density function is

$$P(p) = C p^{b_o} \ln(1/p)^{r_o}$$

with parameters b_o and r_o. When test data consists of $\hat{T}$ operating hours after r failures,

$$\hat{T} = t_1 + t_2 + \ldots + t_r + (m - r)t_r.$$

Here t_r is the time of the r-th failure among m initially on test. Failures are not replaced and the test is terminated at the r-th failure. The resulting posterior density function of p is

$$f(p|\hat{a},\hat{b};a_o,b_o) = \frac{(b+1)^{a+1}}{\Gamma(a+1)} p^b (\ln 1/p)^a,$$

$$0 \leqslant p \leqslant 1,$$

where $a = r + r_o$ and $b = \hat{T}/t + b_o$. The k-th moment of $f(p)$ is

$$M_k\{f\} = (b+1)^{a+1} (k+b+1)^{-a-1}.$$

The above results are from Reference [25].

Component Having Fixed Probability of Success

A single component has an unknown fixed probability of success, p. In testing, there were observed m successes in n trials. For the natural conjugate Beta prior density function with parameters m_o and n_o the posterior density function of p is

$$f(p|a,b) = \frac{1}{B(a+1,b+1)} p^a(1-p)^b$$

where

$$a = m + m_o,\ b = n + n_o - a \text{ and } B(a+1,\ b+1) = \int_o^1 p^a(1-p)^b\, dp.$$

The k-th moment of $f(p|a,b)$,

$k = 0, 1, 2, \ldots$ is:

$$M_k\{f\} = \frac{(b-a+1)!}{a!} \frac{(a+k)!}{(b-a+k)!} = \frac{\Gamma(b-a+2)}{\Gamma(a+1)} \cdot \frac{\Gamma(a+k+1)}{\Gamma(b-a+k+1)}.$$

This result is from [26].

Steady State Availability of Component With Repair

A two state component has exponential distributions of life and of repair times. The duration of intervals of operation and repair define two different statistically independent sequences of identically distributed, mutually independent random variables. Both the mean-up time, $1/\lambda$, and mean repair time $1/\mu$ are unknown parameters estimated from test and prior data.

The long term availability of the component is a function of the random variables μ and λ i.e.:

$$a = \mu/(\lambda + \mu).$$

Assuming gamma priors for λ and μ with snapshot, life and repair time data, the posterior density of availability a is the Euler density function:

$$f(a|r,w,\delta) = \frac{(1-\delta)^w}{B(r,w)} \frac{a^{w-1}(1-a)^{r-1}}{(1-\delta a)^{r+w}},$$

$$0 \leqslant a \leqslant 1;\ r > 0;\ w > 0,\ |\delta| < 1.$$

The parameters r, w and δ are determined by test data and prior information as defined in [25].

The moments of $f(a)$ are given in [25] in terms of Gauss' hypergeometric function ${}_2F_1(w + r,\ w + k;\ w + r + k;\ \delta)$. (Note the typographical error in [25] where k in ${}_2F_1$ is replaced by r.)

A special case of this availability model treating only "snapshot" data is given in [28]. Snapshot data defined in [25,28] records only the state of the system (up or down) at random instants of time.

RULES OF COMBINATION FOR SOME BASIC SYSTEM ELEMENTS

Components are often combined to form system elements which are special in some sense. For example, the same multicomponent element may appear several times as a unit in the same system. In this case, it may be convenient to treat the element as a single system component. Some simple multicomponent system elements are presented in the following:

N Identical Components in Series

The reliability, p, of N identical components in series is $p = p_1^N$.

Component reliability p_1 is a random variable in the Bayes representation with known posterior density, $f_1(p_1)$. The moments $M_{k1}\{f_1\};\ k = 0, 1, \ldots;$ of $f_1(p_1)$ are then also known. The moments $M_k\{f\}$ of the posterior density $f(p)$ of p are related to moments of the f_i by

$$M_k\{f\} = M_{Nk,1}\{f_1\};\ k = 0, 1, 2, \ldots .$$

Using this result one can write the moments of the posterior density of series combinations of any of the special components treated in the previous section.

N Identical Redundant Components

When only one is required to operate in order that the system operates, then the reliability, p, of N identical failure independent redundant components is $p = 1 - (1 - p_1)^N$ where p_1 is the Bayes representation of the component reliability p_1. It is shown in [8] that the moments $M_k\{f\}$ of the posterior density $f(p)$ of p are related to the moments $M_{k1}\{f_1\}$ of the posterior density $f_1(p_1)$ of p_1 by the relation

$$M_k\{f\} = \sum_{j=0}^{k} (-1)^j \binom{k}{j} M_{j1}\{f_1\} .$$

By alternately applying this result and the previous one for components in series, the moments of the posterior density of any series parallel system of components can be obtained.

A "2 out of 3" Element

An element consisting of three identical failure independent components, which operates if any two or more of the components operate, is sometimes called a "2 out of 3 voter," [21]. The structure function of this element is

$$\phi(x_1, x_2, x_3) = 1 \text{ if } x_1 + x_2 + x_3 \geqslant 2$$
$$= 0 \text{ if } x_1 + x_2 + x_3 < 2$$

and the reliability p is

$$p = 3p_1^2 - 2p_1^3$$

where p_1 is the component reliability. If the posterior density $f_1(p_1)$ of p_1 has moments $M_{k1}\{f_1\}$ then the moments $M_k\{f\}$ of the posterior density $f(p)$ of p are:

$$M_k\{f\} = 3^k \sum_{j=0}^{k} \left(-\frac{2}{3}\right)^j \binom{k}{j} M_{j+2k,1}\{f_1\}.$$

This result follows using the fact that for $p = p_1^N$, $M_p\{f(p)\} = M_{Nk,1}\{f_1\}$, when applied term by term to the expansion of

$$(3p_1^2 - 2p_1^3)^k.$$

Reference [21] gives the reliability function of the N-tuple Modular Redundant design consisting of N replicated units feeding a $(n + 1)$-out-of-N voter. This case can also be treated by the present methods.

Exactly L Out of N Element

An element consisting of N identical failure independent components which operates only when exactly L out of N components operate is a rather unusual system. If $L + 1$ out of N operate the system fails. Such a system is not a coherent structure in the sense of [5]. The reliability p of this element is given by

$$p = \binom{N}{L} p_1^L (1 - p_1)^{N-L}.$$

The moments of the posterior density $f(p)$ of the Bayes representation p in terms of the moments $M_{k,1}\{f_1\}$ of the posterior density $f_1(p_1)$ of the component reliability p_1 can be shown to be

$$M_k\{f\} = \binom{N}{L}^k \sum_{j=0}^{(N-L)k} (-1)^j \binom{(N-L)k}{j} M_{j+kL,1}\{f_1\}.$$

This example serves to illustrate that the proposed evaluation is not restricted to coherent systems.

DEVELOPMENT OF AN APPROXIMATE PRIOR FOR TESTING AT SYSTEM LEVEL

Section 9.4.4 of NAVORD OD 44622, Reference [22], presents a procedure for developing the posterior beta distribution of system reliability for system level TECHEVAL/OPEVAL testing. Reference [9] presents further discussion with an example. The observed system level data is binomial i.e., r failures in n trials. The system level, natural conjugate prior is the beta density. An exact prior for the system level tests is the posterior density function based on all prior component tests and component priors and can be computed by the methods above. The procedure recommended in OD44622 is to approximate the exact system prior with a beta density having the same first and second moments.

Equation (15) above provides a tractable tool for computing the required first and second moments for extending the method to arbitrary system structures.

Let M_1 and M_2 denote the first and second moments computed as shown in this report for the posterior density $f(R)$ of system reliability, R, based on prior component data. The

$f(R)$ is considered the exact prior for determinations of a new posterior density based on binomial system level data. What are required for the approximation are the parameters n' and r' of the beta prior

$$g(R|n';\ r') = \frac{R^{n'-r'}(1-R)^{r'}}{B(n'-r'+1,\ r'+1)}$$

with the same first and second moments as $f(R)$. Having computed M_1 and M_2 the answer is direct using formulas on page 9.23 of NAVORD OD 44622 i.e.,

$$n' = [M_1(1-M_1)/(M_2-M_1^2)]-1$$
$$r' = (1-M_1)n'.$$

The gamma prior is treated in a similar way in the same reference.

The beta approximation can also be used directly as an approximation to the exact posterior density function for complex systems based on component test data. The approximation has been very good when compared with the exact result in examples treated by the authors. The calculation is tractable for hand computation since only the first and second moments of the exact posterior density function are required.

Numerical Example

Consider a system consisting of five components, A_i $(i = 1, \ldots, 5)$ connected in series. Components A_1, A_2, A_3, and A_4 have unknown fixed probabilities of success, p_i; and in testing, there were observed m_i successes in n_i trials. The fifth component, A_5, has an unknown constant failure rate λ and has mission time t. In testing, component A_5 failed r times in T operating hours. The following test data were observed:

$$n_1=20,\ m_1=18;\ n_2=30,\ m_2=25;\ n_3=20,\ m_3=20;\ n_4=20,\ m_4=19;\ T=38,\ t=6,\ r=3$$

The resulting posterior density functions are:

$$f_1(R_1) = 3990\ R_1^{18}\ (1-R_1)^2$$
$$f_2(R_2) = 4417686\ R_2^{25}\ (1-R_2)^5$$
$$f_3(R_3) = 21\ R_3^{20}$$
$$f_4(R_4) = 420\ R_4^{19}\ (1-R_4)$$
$$f_5(R_5) = 482.00823\ R_5^{19/3}\left(\ln\frac{1}{R_5}\right)^3.$$

We know [25,26] that the Mellin integral transform of the posterior density function, $h(R)$ for the system is the product of the Mellin integral transforms of the density functions of the components. At this point we can determine $h(R)$ exactly by means of the inverse Mellin integral transform or we can approximate $h(R)$ with a Beta density function having the same first and second moments as $h(R)$.

The Mellin integral transforms of the density function for the components of the system are:

$$M[f_1(R_1)|S] = \frac{21!}{18!}\frac{\Gamma(S+18)}{\Gamma(S+21)} \qquad M[f_4(R_4)|S] = \frac{21!}{19!}\frac{\Gamma(S+19)}{\Gamma(S+21)}$$

$$M[f_2(R_2)|S] = \frac{31!}{25!}\frac{\Gamma(S+25)}{\Gamma(S+31)} \qquad M[f_5(R_5)|S] = \frac{(22/3)^4}{(S+19/3)^4}$$

$$M[f_3(R_3)|S] = \frac{21!}{20!}\frac{\Gamma(S+20)}{\Gamma(S+21)}$$

The Mellin Integral transform of $h(R)$ is $M[h(R)|S] = \prod_{i=1}^{5} M[f_i(R_i)|S]$.

From [26] we know that the Mellin inversion integral yields directly

$$h(R) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} R^{-S} M[h(R)|S]\,dS$$

where the path of integration is any line parallel to the imaginary axis and lying to the right of the real part of c, If b is greater than 1, the real part of c is greater than p, and p is any number, then, [26]

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{R^{-S}}{(S+p)^b}\,dS = \frac{R^p}{\Gamma(b)}\left[\ln\left(\frac{1}{R}\right)\right]^{b-1}.$$

To find $h(R)$ we simply write $M[h(R)|S]$ as the sum of its partial fractions [13] and integrate each term using the above equation. Thus the exact posterior density function, $h(R)$, for system reliability is

$$\begin{aligned} h(R) = {} & + 1094388844.948\ R^{18} + 30505643166.29\ R^{19}. \\ & - 12601708553.76\ R^{19}\left(\ln\frac{1}{R}\right)^1 - 31650550963.66\ R^{20} \\ & - 19915799047.82\ R^{20}\left(\ln\frac{1}{R}\right) - 5114357474.61\ R^{20}\left(\ln\frac{1}{R}\right)^2 \\ & + 235122603.404\ R^{25} - 354959810.01\ R^{26} \\ & + 249501799.456\ R^{27} - 98389473.63\ R^{28} \\ & + 21240815.37\ R^{29} - 1974044.939\ R^{30} \\ & - 22937.221\ R^{19/3} + 78073.717\ R^{19/3}\left(\ln\frac{1}{R}\right) \\ & - 95839.296\ R^{19/3}\left(\ln\frac{1}{R}\right)^2 + 42683.275\ R^{19/3}\left(\ln\frac{1}{R}\right)^3 \end{aligned}$$

The exact distribution function, $H(R)$, is found by integrating the density function.

To obtain the approximate solutions for the system reliability density and distribution functions, we recall that the first and second moments of $h(R)$ are given by $M[h(R)|2]$ and $M[h(R)|3]$ respectively. The beta density function, which is used to approximate $h(R)$, is

$$\hat{h}(R) = \frac{R^a(1-R)^b}{B(a+1,\ b+1)}$$

where $\hat{h}(R)$ denotes the approximate system density function, $\beta(a+1,\ b+1)$ is the complete beta function, and a and b are the parameters of the beta function. The first moment of $\hat{h}(R)$ is

$$\frac{a+1}{a+b+2}$$

and the second moment is

$$\frac{(a+1)\,(a+2)}{(a+b+2)\,(a+b+3)}\,.$$

We require that the first and second moments of $h(R)$ and $\hat{h}(R)$ be equal. Thus we have

$$M[h(R)|2] = \frac{a+1}{a+b+2} =$$

$$M[h(R)|3] = \frac{(a+1)\,(a+2)}{(a+b+2)\,(a+b+3)}\,.$$

Solving simultaneously for a and b yields the parameters for the beta density function. Thus we have $a = 6.43596$ and $b = 11.92734$. Therefore we can now write $\hat{h}(R)$, the approximate density function for system reliability:

$$\hat{h}(R) = \frac{R^{6.43596}\,(1-R)^{11.92734}}{B\,(7.43596,\ 12.92734)}.$$

To determine the approximate distribution functions, $H(R)$, for system reliability we simply integrate $h(R)$

Table 1 provides the comparison between the results obtained by the exact solution and the approximate solution.

TABLE 1 — *Numerical Results Obtained from Exact and Approximate Solutions*

R	Density Function		Distribution Function	
	Exact $h(R)$	Approximate $\hat{h}(R')$	Exact $H(R)$	Approximate $\hat{H}(R)$
.0	.0	.0	.0	.0
.10	.079	.057	.001	.001
.20	1.213	1.208	.052	.048
.30	3.243	3.339	.278	.281
.40	3.429	3.382	.635	.641
.50	1.667	1.616	.896	.895
.60	.343	.365	.986	.984
.70	.020	.032	.999	.999
.80	.003	.001	1.000	1.000
.90	0.000	0.000	1.000	1.000
1.00	0.000	0.000	1.000	1.000

REFERENCES

[1] Abramowitz, M. and I.A. Stegun (Editors), "Handbook of Mathematical Functions," National Bureau of Standards, Applied Mathematics Series, *55*, 782 (1964).

[2] Aitchison, J., "Two Papers on the Comparison of Bayesian and Frequentist Approaches to Statistical Problems of Prediction," Journal of the Royal Society, Series B., *26*, 161-175 (1964).

[3] Bartholomew, D.J., "A Comparison of Some Bayesian and Frequentist Inferences," Biometrika *52*, 1 and 2, 19-35 (1965).

[4] Birnbaum, Z.W., "On the Probabilistic Theory of Complex Structures," Proceeding of the Fourth Berkeley Symposium, *1*, 49-55, University of California Press (1961).

[5] Birnbaum, Z.W., J.D. Esary and S.C. Saunders, "Multi-Component Systems and Structures and Their Reliability," Technometrics, *3*, 55-77 (1961).

[6] Brender, D.M., "Reliability Testing in a Bayesian Context," IEEE International Convention Record, Part 9, 125-136 (1966).

[7] Chang, E.Y. and W.E. Thompson, "Bayes Analysis of Reliability of Complex Systems," Operations Research, *24*, 156-168 (1976).

[8] Chang, E.Y. and W.E. Thompson, "Bayes Confidence Limits for Reliability of Redundant Systems," Technometrics, *17* (1975).

[9] Cole, Peter Z.V., "A Bayesian Reliability Assessment of Complex Systems for Binomial Sampling," IEEE Transactions on Reliability, *R-24*, 114-117 (1975).

[10] Esary, J.D. and F. Proschan, "Coherent Structures of Non-Identical Components," Technometrics, *5*, 191-209 (1963).

[11] Esary, J.D. and F. Proschan, "The Reliability of Coherent Systems," *Redundancy Techniques for Computing Systems*, Edited by R.H. Wilcox and W.C. Mann, SPARTAN BOOKS, 47-61 (1962).

[12] Fox, B.L., "A Bayesian Approach to Reliability Assessment," NASA Memorandum RM/5084 - (1966).

[13] Gardner, M.F. and J.L. Barnes, *Transients In Linear Systems*, John Wiley and Sons, New York, 152-163 (1942).

[14] Gaver, D.P. and M. Mazumdar, "Statistical Estimation in a Problem of System Reliability," Naval Research Logistics Quarterly, *14*, 473-488 (1967).

[15] Gnedenko, B.V., Yu.K. Belyayev and A.D. Solovyev, *Mathematical Methods of Reliability Theory*, Academic Press, New York, 76-77 (1969).

[16] Lanczos, C., *Applied Analysis*, Prentice Hall, Inc. Chapter IV and VII.

[17] Lin, P.M., B.J. Leon and T.C. Huang, "A New Algorithm for Symbolic System Reliability Analysis," IEEE Transactions on Reliability, *R-25* (1976).

[18] Lindley, D.V., "The Robustness of Internal Estimates," Bulletin of International Statistical Institute, *38*, 209-220 (1961).

[19] Lindley, D.V., "The Use of Prior Probability Distributions in Statistical Inference and Decisions," Proceeding of the Fourth Berkeley Symposium, *1*, 453-468, University of California Press (1961).

[20] Maritz, J.S., "Empirical Bayes Methods," Methuen's Monographs on Applied Probability and Statistics, Methuen and Co. Ltd., London (1970).

[21] Matther, F.P. and Paulo T. deSousa, "Reliability Models of N-tape Modular Redundancy Systems," IEEE Transactions on Reliability, *R-24*, 108 (1975).

[22] NAVORD OD 44622, Reliability Guide Series, *4*. The Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

[23] Parker, J.B., "Bayesian Prior Distributions for Multi-Component Systems," Naval Research Logistics Quarterly, *19* (1972).

[24] Savage, L.J., "The Foundations of Statistics Reconsidered," Proceeding of the Fourth Berkeley Symposium, *1*, 575-585, University of California Press (1961).

[25] Springer, M.D. and W.E. Thompson, "Bayesian Confidence Limits for the Reliability of Cascade Exponential Subsystems," IEEE Transactions on Reliability, *R-16*, 86-89 (1967).

[26] Springer, M.D. and W.E. Thompson, "Bayesian Confidence Limits for the Product of N Binomial Parameters," Biometrika *53*, 3 and 4, 611 (1966).

[27] Thompson, W.E. and P.A. Palicio, "Bayesian Confidence Limits for the Availability of Systems," IEEE Transactions on Reliability, *R-24*, 118-120 (1975).

[28] Thompson, W.E. and M.D. Springer, "A Bayes Analysis of Availability for a System Consisting of Several Independent Subsystems," IEEE Transactions on Reliability, *R-21*, 212-214 (1972).

[29] Wolf, J.E., "Bayesian Reliability Assessment From Test Data," Proceedings 1976 Annual Reliability and Maintainability Symposium, Las Vegas, Nevada, *20-22*, 411-419 (1976).

OPTIMAL REPLACEMENT OF PARTS HAVING OBSERVABLE CORRELATED STAGES OF DETERIORATION*

L. Shaw

Polytechnic Institute of New York
Brooklyn, New York

C-L. Hsu

Minneapolis Honeywell
Minneapolis, Minnesota

S. G. Tyan

M/A COM Laboratories
Germantown, Maryland

ABSTRACT

A single component system is assumed to progress through a finite number of increasingly bad levels of deterioration. The system with level i $(0 \leqslant i \leqslant n)$ starts in state 0 when new, and is definitely replaced upon reaching the worthless state n. It is assumed that the transition times are directly monitored and the admissible class of strategies allows substitution of a new component only at such transition times. The durations in various deterioration levels are dependent random variables with exponential marginal distributions and a particularly convenient joint distribution. Strategies are chosen to maximize the average rewards per unit time. For some reward functions (with the reward rate depending on the state and the duration in this state) the knowledge of previous state duration provides useful information about the rate of deterioration.

Many authors have studied optimal replacement rules for parts characterized by Markovian deterioration, for example Kao [6] and Luss [9] and the many references found in those papers. Kao minimized the expected average cost per unit time for semi-Markovian deteriorating system, and considered various combinations of state and age-dependent replacement rules.

Luss examined inspection and repair models, where he assumed that the operating costs occurring during the system's life increase with the increasing deterioration. The holding times in the various states were independently, identically, and exponentially distributed. The policies examined include the scheduling of the next inspections (when an inspection reveals that the state of the system is better than certain critical state k) and preventive repairs (when an inspection reveals the state of the system being worse than or equal to k). The convenience of a Poisson-type structure for the number of events-per-unit-time made it relatively easy to allow general freedom in the selection of observation times.

The work studied here is based on a modification of the model used by Luss. Our model for deterioration is more general, but the admissible strategies used here are more restricted. Here we allow the exponentially distributed durations to have different mean values, and to be positively correlated.

*This work was partially supported by Grant No. N00014-75-C-0858 from the Office of Naval Research.

The introduction here of correlation between interval durations permits the modeling of a rate of deterioration which can be estimated from a particular realization of the past durations. However, the lack of a Poisson-type of structure for the events-per-unit-time makes it much more difficult here to allow general freedom in the selection of observation times. At present, only the simple case of direct and instantaneous observation of deterioration jumps has been considered.

This model would be appropriate, for example, in a subsystem which functions, but with reduced efficiency, when some redundant components have failed; and for which failure of one component might indicate environmental stresses which increase the probability of failure for other components. In addition, deterioration in correlated stages might be used as a simple approximation for a continuously varying degradation which does not exhibit discrete stages.

Figure 1 shows a typical time history of deterioration and replacement. The duration in state $(i-1)$, prior to reaching state (i), is r_{i-1}. The intervals d_i in Figure 1 represent the time required to replace a component when it has entered state i. The sequence $\{r_i\}$ will be Markov, characterized by a multi-variate exponential distribution. Reward functions will be related to the deterioration state and the time spent in each state. The decision rule specifies whether or not to replace when entering each state i, on the basis of the history of $r_{i-1}, r_{i-2}, \ldots$. The Markov property simplifies the decision rule to be a collection of $\mathcal{C}_i$ sets such that we replace on entering state i if and only if $r_{i-1} \in \mathcal{C}_i$.

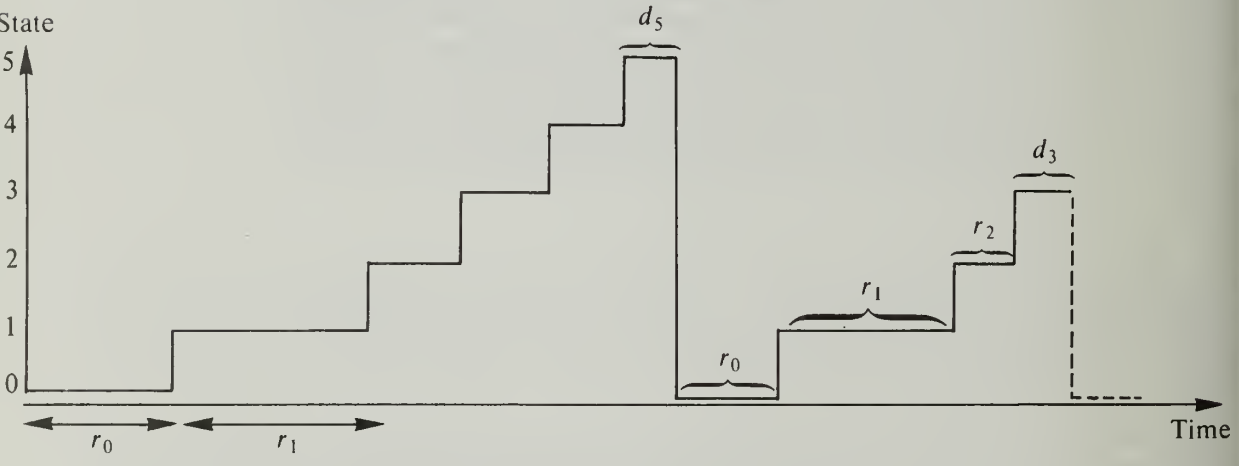


FIGURE 1. History of deterioration and replacement ($n = 5$).

The objective is to maximize the average reward per unit time:

$$L = \lim_{T\to\infty} \frac{1}{T} \text{ (Total reward in } (0,T)) \tag{1}$$

$$= \frac{E[\text{Reward per renewal}]}{E[\text{Duration between renewals}]} = \frac{\mathcal{R}}{\mathcal{D}}. \tag{2}$$

(See Ross [11] page 160 for equivalence of (1) and (2).) The mean reward per renewal is defined here as:

$$\mathcal{R} = E\left[\sum_{i=0}^{N-1} \int_0^{r_i} c_i(t)\,dt - p_N\right], \tag{3}$$

in which:

N = state at which replacement occurs (possibly random).

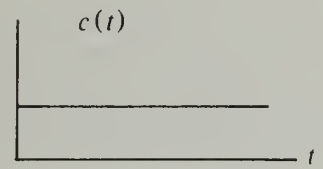
p_N = replacement cost if replaced on entering state N (possibly random).

$c_i(t)$ = reward rate when in state i.

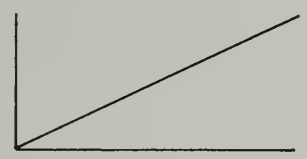
Figure 2 shows several reward rate time functions $c(t)$ which have been considered. When one of these $c(t)$ functions is specified for a given problem, the $c_i(t)$ in (3) are assigned values $\beta_i c(t)$ with:

$$\beta_0 \geqslant \beta_1 \geqslant \beta_2 \geqslant \ldots \geqslant \beta_{n-1} \geqslant \beta_n = 0, \tag{4}$$

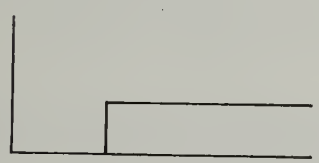
to assure greater reward rates in less deteriorated states. State n corresponds to a completely failed or worthless component.



(a) constant



(b) linear



(c) constant-after set up

FIGURE 2. Reward rate time functions.

The mean duration in (2) is defined as:

$$\mathfrak{D} = E\left[\sum_{i=0}^{N-1} r_i + d_N\right], \tag{5}$$

to include a possibly random time d_N for carrying out a replacement at state N.

While the ultimate objective is to choose $\mathcal{C}_i$ to maximize the L defined in (1), it is well known that a related problem of maximizing:

$$\mathfrak{L}_0(\alpha) = \mathfrak{R} - \alpha\, \mathfrak{D}, \tag{6}$$

is simpler [1]. Indeed, the $\mathcal{C}_i$ which maximize L will be identical to those which maximize $\mathfrak{L}_0(\alpha)$ for the α^* such that:

$$\mathfrak{L}_0^0(\alpha^*) = 0, \text{ where } \mathfrak{L}_0^0(\alpha) \triangleq \max_{\{\mathcal{C}_i\}} \mathfrak{L}_0(\alpha). \tag{7}$$

Section 1 considers a case in which it is found that deterioration rate information is not useful (e.g., the optimal policy is independent of the amount of correlation between successive state durations).

Sections 2 and 3 consider other penalty cost structures, e.g., assuming that more deteriorated parts are rustier, hotter, or more brittle, and therefore more costly to replace. In such cases the optimal policies do make use of estimates of the deterioration rates as well as of observations of the deterioration level.

The Appendix describes useful properties of the multivariate exponential $\{r_i\}$ sequence which is used to model the correlated residence times in a sequence of deterioration states. These durations have marginal distributions which are exponential with mean values η_i, and correlations $\rho_{r_i r_j} = \rho^{|i-j|}$.

1. CONSTANT REWARD RATE—STATE INDEPENDENT REPLACEMENT PENALTIES

The constant reward rate case with $c_i(t) = \beta_i$ and with state-independent replacement penalties ($p_i = p$, $d_i = d$) is particularly simple to analyze. We will see that as long as $E[r_i | r_{i-1}, r_{i-2}, \ldots] \geqslant 0$ for all i, even if the r_i are not exponentially distributed, the optimal rule will be to replace the deteriorating part upon entering some critical state k^*, independent of the observed durations r_i.

Based on the problem statement, the optimal decision on entering state j must maximize the mean future reward until the next renewal, $\mathfrak{L}_j(\alpha)$, for a suitable α. Here:

$$\mathfrak{L}_j(\alpha) = E\left[\sum_{i=j}^{N-1} \beta_i r_i | r_{j-1}\right] - \alpha E\left[\sum_{i=j}^{N-1} r_i | r_{j-1}\right] - p - \alpha d. \tag{8}$$

Immediately after a renewal, when $j = 0$, the expectations defining $\mathfrak{L}_0(\alpha)$ are unconditional. The optimal decisions for each state will be found in terms of α, and then the proper α^* (for producing decisions which maximize L) is the one for which the maximum:

$$\max \mathfrak{L}_0(\alpha^*) = \mathfrak{L}_0^0(\alpha^*) = 0. \tag{9}$$

Optimization by dynamic programming begins by considering the decisions at the last step, i.e., on entering state $(n-1)$. There are two choices, to replace (R) or not to replace $(\bar{R})$, with corresponding values:

$$\mathfrak{L}_{n-1}(\alpha; R) = -p - \alpha d, \tag{10}$$

and:

$$\begin{aligned} \mathfrak{L}_{n-1}(\alpha; \bar{R}) &= E[\beta_{n-1} r_{n-1} | r_{n-2}] - \alpha E[r_{n-1} | r_{n-2}] - p - \alpha d \\ &= E[(\beta_{n-1} - \alpha) r_{n-1} | r_{n-2}] - p - \alpha d. \end{aligned} \tag{11}$$

Clearly, the best decision is not to replace, if and only if, the difference

$$\begin{aligned} \Delta_{n-1}(\alpha; r_{n-2}) &\triangleq \mathfrak{L}_{n-1}(\alpha; \bar{R}) - \mathfrak{L}_{n-1}(\alpha; R) \\ &= (\beta_{n-1} - \alpha) E[r_{n-1} | r_{n-2}] \geqslant 0. \end{aligned} \tag{12}$$

is non-negative. The sign of (12) will be the sign of $(\beta_{n-1} - \alpha)$, due to the non-negativity of all interval durations. Thus the best decision depends on α and the reward parameter β_{n-1}, but not on the previously observed duration. Two cases will be considered separately.

If $\beta_{n-1} \geqslant \alpha$ then the best decision at state $(n-1)$ is not to replace. We will now explain why, under this condition, it is best not to replace at any state less than n. Consider the situation on entering $(n-2)$. We have already shown that it is better not to replace on entering $(n-1)$. Thus the choice will be based on a Δ_{n-2} of the form:

$$\Delta_{n-2}(\alpha;r_{n-3}) = E[(\beta_{n-2}-\alpha)r_{n-2} + (\beta_{n-1}-\alpha)r_{n-1}|r_{n-3}). \tag{13}$$

Here we have:

$$(\beta_{n-2}-\alpha) > (\beta_{n-1}-\alpha) > 0, \tag{14}$$

by assumption, and:

$$E[r_{n-1}|r_{n-3}] \geqslant 0 \text{ and } E[r_{n-2}|r_{n-3}] \geqslant 0, \tag{15}$$

because all $r_i \geqslant 0$ with probability one. Thus $\Delta_{n-2}(\alpha;r_{n-3}) > 0$ for all $r_{n-3} > 0$, and it is also better not to replace here. This argument can be repeated for states $(n-3)$, $(n-4)$, ..., 1, 0.

The other case to consider is $\beta_{n-1} < \alpha$, which requires replacement on entering state $(n-1)$, if the system ever reaches that state. When we consider the decision on entering $(n-2)$, the Δ_{n-2} is:

$$\Delta_{n-2}(\alpha;r_{n-3}) = E[(\beta_{n-2}-\alpha)r_{n-2}|r_{n-3}], \tag{16}$$

which has the sign of $(\beta_{n-2}-\alpha)$. If $(\beta_{n-2}-\alpha) < 0$, then replacement is optimal on entering $(n-2)$ and $(n-3)$ is considered next. This iteration may eventually reach a state $(k-1)$ where $(\beta_{k-1}-\alpha) > 0$ and it is better not to replace. Arguments similar to those for the $\beta_{n-1}-\alpha > 0$ case show that nonreplacement is the optimal decision at all states preceding the one which first arises as a nonreplacement state in this backward iteration.

In summary, in the constant reward rate-constant replacement penalty case $\mathfrak{L}_0(\alpha)$ is maximized by a decision rule which says replace on entering some state $k \leqslant n$ which depends on the reward parameters $\{\beta_i\}$ and the α:

$$k = \min\{i:(\alpha - \beta_i) > 0\}. \tag{17}$$

Finally, we must choose α^* so that $\mathfrak{L}_0^0(\alpha^*) = 0$, where:

$$\mathfrak{L}_0^0(\alpha) = -p - \alpha d + \sum_0^{k-1} (\beta_i - \alpha)E[r_i]. \tag{18}$$

Figure 3 shows a typical plot of $\mathfrak{L}_0^0(\alpha)$ as a continuous, piecewise linear curve whose zero crossing ($\mathfrak{L}_0^0(\alpha^*) = 0$) defines α^* and the optimal replacement state k^* for maximizing L.

EXAMPLE: Figure 3 shows that the optimal average reward per unit time is $L = 2\frac{5}{7}$ when $k^* = 3$, where $\beta_0 = 5, \beta_1 = 4, \beta_2 = 3, \beta_3 = 2, \beta_4 = 1, \beta_5 = 0, p = 5, d = 1, \eta_i = 2$ $(i = 0,1,2,3,4)$ and $n = 5$. From Equation (18), the optimal k is a function of α, which remains constant when α varies over each interval $\beta_{i+1} < \alpha \leqslant \beta_i$, as shown in the figure.

2. INCREASING REPLACEMENT PENALTIES-CONSTANT REWARD RATE

Here we generalize the model of the previous section by allowing the replacement cost p_i and replacement duration d_i to be functions of the replacement state (i), and to be random. These parameters are assumed to have mean values $E[p_i]$ and $E[d_i]$ which are convex nondecreasing sequences in i, corresponding to the increased difficulty in replacing more deteriorated parts which may be, e.g., rustier, hotter or more brittle. We also assume that the mean durations are ordered: $\eta_0 \geqslant \eta_1 \geqslant \ldots \geqslant \eta_{n-1}$, corresponding to faster transitions of more deteriorated parts.

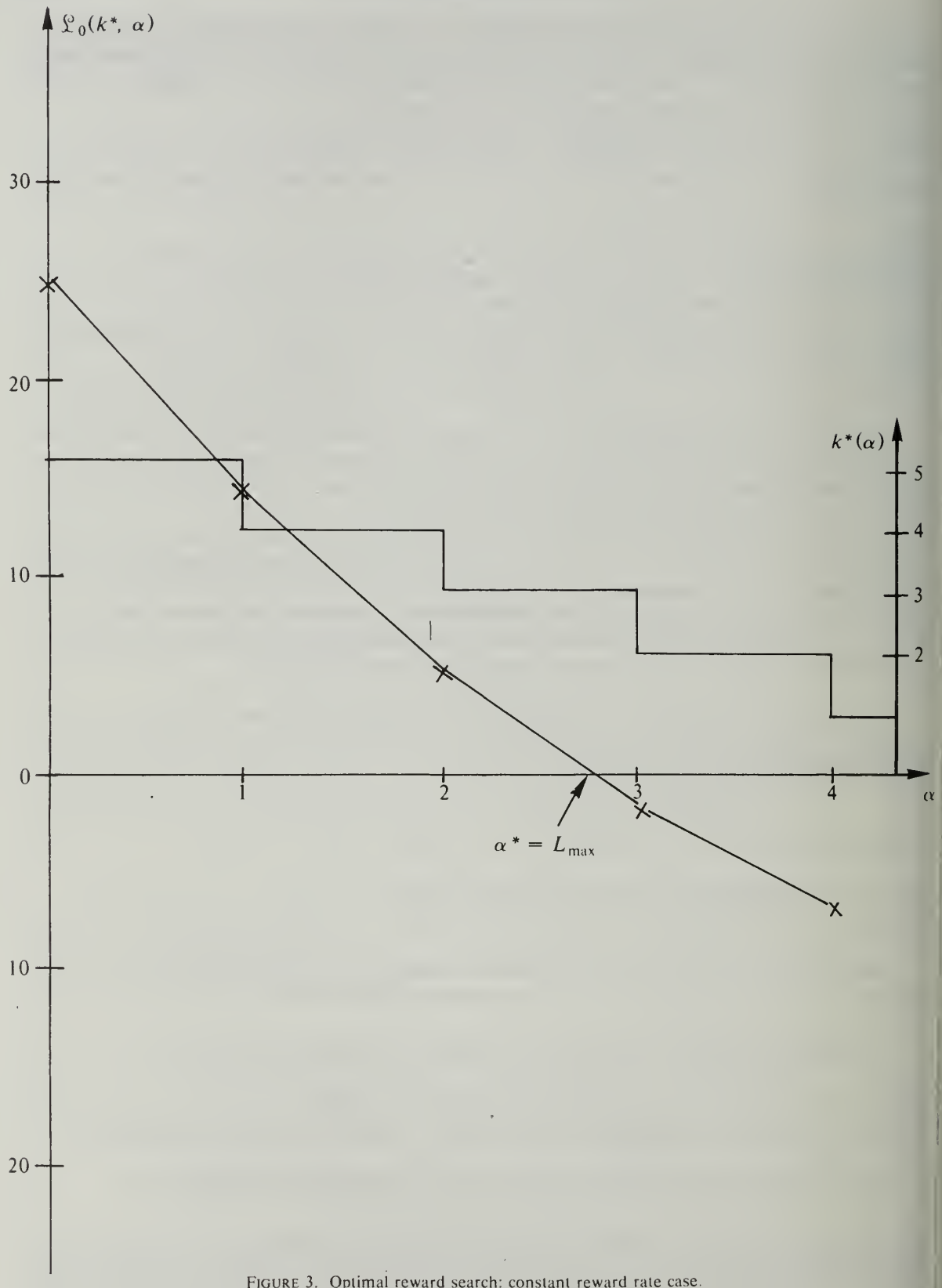


FIGURE 3. Optimal reward search: constant reward rate case.

The foregoing assumptions, together with properties of the assumed multivariate exponential density for stage-durations (see Appendix), lead to an optimal decision policy with a nice structure. That optimal policy prescribes replacement when entering state j, if and only if $r_{j-i} < r^*_{j-1}$, where the decision thresholds are ordered: $0 \leqslant r^*_0/\eta_0 \leqslant r^*_1/\eta_1 \leqslant \ldots \leqslant r^*_{n-1}/\eta_{n-1} = \infty$.

The optimal decision on entering state j must maximize the mean future reward until the next renewal, i. e., $\mathfrak{L}_j(\alpha)$. For a suitable α, we have:

$$\mathfrak{L}_j(\alpha) = E\left\{\sum_{i=j}^{N-1} \beta_i\, r_i | r_{j-1}\right\} - \alpha E\left\{\sum_{i=j}^{N-1} r_i | r_{j-1}\right\} - E[p_N + \alpha d_N]. \tag{19}$$

For notational simplicity we define $e_i = E[p_i + \alpha d_i]$ and note that e_i is also convex and nondecreasing since we are only interested in $\alpha > 0$. The optimal decisions for each state will be found in terms of α, and then the proper α^* (for producing decisions which maximize L) is the one for which the maximum $\mathfrak{L}$ vanishes:

$$\mathfrak{L}^0_0(\alpha^*) = -\, e_N(\alpha^*) + E\left[\sum_{i=0}^{N-1} \beta_i\, r_i - \alpha^* \sum_{i=0}^{N-1} r_i\right] = 0. \tag{20}$$

Optimization by dynamic programming begins by considering the decision at the last step. Since state n represents a failed component, we definitely replace the component when it enters state n. Next, we consider the decision to be made on entering state $n-1$. There are two choices: to replace (R) or not to replace $(\bar{R})$, with corresponding values

$$\mathfrak{L}_{n-1}(\alpha;R) = -e_{n-1}, \tag{21}$$

$$\mathfrak{L}_{n-1}(\alpha;\bar{R}) = E[\beta_{n-1} r_{n-1} - \alpha r_{n-1} | r_{n-2}] - e_n \tag{22}$$

for $\mathfrak{L}_{n-1}(\alpha)$. Clearly, the best decision is not to replace, if and only if,

$$\Delta_{n-1}(r_{n-2}) \triangleq \mathfrak{L}_{n-1}(\alpha;\bar{R}) - \mathfrak{L}_{n-1}(\alpha,R)$$

is non-negative, i.e.,

$$\Delta_{n-1}(r_{n-2}) = (\beta_{n-1} - \alpha)\, E[r_{n-1} | r_{n-2}] + (e_{n-1} - e_n) \geqslant 0. \tag{23}$$

Referring to (A-6), $\Delta_{n-1}(r_{n-2})$ is a linear function of r_{n-2}, with

$$\Delta_{n-1}(0) = (\beta_{n-1} - \alpha)\eta_{n-1}(1-\rho) + (e_{n-1} - e_n).$$

Figure 4 shows the possible shapes for this function. There can be no downward zero-crossing at an $r_{n-2} > 0$.

Thus, depending on the numerical values of the parameters, there are three possible kinds of optimal decision rules when entering state $(n-1)$:

(i) replace for no r_{n-2} if $\Delta_{n-1} \geqslant 0$ for all $r_{n-2} \geqslant 0$

(ii) replace for any r_{n-2} if $\Delta_{n-1} \leqslant 0$ for all $r_{n-2} \geqslant 0$

(iii) replace if and only if $r^*_{n-2} > r_{n-2} \geqslant 0$, where $\Delta_{n-1}(r^*_{n-2}) = 0$.

In other words,

$$\mathcal{C}_{n-1}(\alpha) = \{r_{n-2}\colon r_{n-2} < r^*_{n-2}\}, \tag{24}$$

where r^*_{n-2} could be zero (case i) or infinite (case ii).

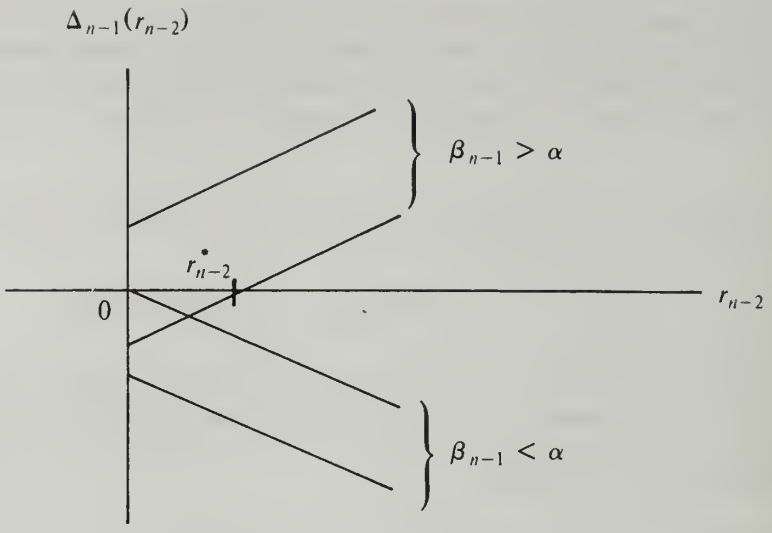


FIGURE 4. Possible shapes for $\Delta_{n-1}(r_{n-2})$.

Next we consider the optimal decision when entering state $(n-2)$, and assuming that the optimal decision will be made at the subsequent stage. We consider cases of $(\beta_{n-1} < \alpha)$ and $(\beta_{n-1} \geqslant \alpha)$ separately.

(a) $(\beta_{n-1} < \alpha)$ implies replacement on entering $(n-1)$, so

$$\Delta_{n-2}(r_{n-3}) = (\beta_{n-2} - \alpha)\, E[r_{n-2}|r_{n-3}] + (e_{n-2} - e_{n-1}),$$

resulting in the same three possibilities listed above for state $(n-1)$.

(b) for $(\beta_{n-1} > \alpha)$:

$$\begin{aligned}\Delta_{n-2}(r_{n-3}) = {} & e_{n-2} + (\beta_{n-2} - \alpha) E[r_{n-2}|r_{n-3}] \\ & + \int_{r^{\bullet}_{n-2}}^{\infty} [(\beta_{n-1} - \alpha)\, E[r_{n-1}|r_{n-2}] - e_n]\, f(r_{n-2}|r_{n-3})\, dr_{n-2} \\ & + \int_{0}^{r^{\bullet}_{n-2}} (-e_{n-1})\, f(r_{n-2}|r_{n-3})\, dr_{n-2}\end{aligned} \tag{25}$$

Equation (25) can be simplified, with the aid of the notation $(x)^+ = \max(x, 0)$, to the form

$$\begin{aligned}\Delta_{n-2}(r_{n-3}) = {} & (e_{n-2} - e_{n-1}) + (\beta_{n-2} - \alpha)\, E[r_{n-2}|r_{n-3}] \\ & + E[(\Delta_{n-1}(r_{n-2}))^+|r_{n-3}].\end{aligned} \tag{26}$$

Useful comparisons can be formed if normalized variables are introduced, namely

$$s_i = r_i/\eta_i;\ \delta_i(s_{i-1}) = \Delta_i(r_{i-1})\big|_{r_{i-1}=\eta_{i-1}s_{i-1}}$$

We now prove

(a) $\delta_{n-2}(s_{n-3}) \geqslant \delta_{n-1}(s_{n-3})$

(b) $\delta_{n-2}(s_{n-3})$ is convex with at most one upward zero crossing at an $s > 0$.

There is no harm in writing $\delta_{n-1}(s_{n-3})$ or $\delta_{n-1}(s_+)$ instead of $\delta_{n-1}(s_{n-2})$ for purposes of comparing functions.

To prove (a), consider

$$\delta_{n-2}(s) - \delta_{n-1}(s) = [(e_{n-2} - e_{n-1}) - (e_{n-1} - e_n)] + E[(\delta_{n-1}(s_+))^+|s] + [(\beta_{n-2}-\alpha)\eta_{n-2} - (\beta_{n-1}-\alpha)\eta_{n-1}]\, E[s_+|s]. \tag{27}$$

where s_+ represents the normalized duration following s.

The terms on the right side of (27) are nonnegative due to the convexity of the e_i, $(\,)^+ \geqslant 0$, (A-6), and the assumed orderings of the β_i and η_i.

This completes the proof that (a) is true. It follows immediately that if (i) (preceding Eq. (24)) applies for state $(n-1)$, then it is also optimal *not* to replace in state $(n-2)$ or any earlier state. (Recall $\beta_{n-1} < \beta_{n-2} < \ldots$, and we are now considering $\alpha < \beta_{n-1}$).

To prove (b), which is only of interest when an $r^*_{n-2} > 0$ exists, we refer to the theorem in the appendix. The test difference $\delta_{n-2}(s)$ can be written as

$$\delta_{n-2}(s) = E[e_{n-2} - e_{n-1} + (\beta_{n-2}-\alpha)\eta_{n-2}\, s_+ + (\delta_{n-1}(s_+))^+|s] \tag{28}$$

in which the integrand has the properties required by $h(s)$ in the theorem. To see this, we note that $r^*_{n-2} > 0$ implies that $(\delta_{n-1}(0))^+ = 0$, so the integrand is nonpositive at $s_+ = 0$. Thus, $\delta_{n-2}(s)$ has the shape stated in (b), implying that

$$\mathcal{C}_{n-3} = \{r_{n-3}\colon r_{n-3} \leqslant r^*_{n-3}\} \tag{29}$$

where r^*_{n-3} may be zero, infinity, or the nonnegative value defined by $\delta_{n-2}(r^*_{n-3}/\eta_{n-3}) = 0$.

The foregoing arguments can be repeated for r_{n-4}, $r_{n-5} \ldots r_0$ to prove that the optimal replacement policy has the form:

Replace on entering state i, if and only if, $r_i \leqslant r_i^*$ where

$$0 \leqslant r_0^*/\eta_0 \leqslant r_1^*/\eta_1 \leqslant \ldots \leqslant r_{n-1}^*/\eta_{n-1} = \infty.$$

When repeating the proof for earlier stages, the $(\,)^+$ term in (27) and (28) is modified to the form, e.g., $[(\delta_{n-2}(s_+))^+ - (\delta_{n-1}(s_+))^+]$. This term is generally nonnegative, due to (a) at the preceding iteration (next time step); and it is zero for $s_+ = 0$ when proving (b), since then $r^*_{n-3} > 0$. Thus the basic theorem is still applicable.

3. Computational Procedure

The preceding section derived the structure of the optimal decision rule for the case where replacement is more difficult and more expensive when the part is more deteriorated. The corresponding optimal decision thresholds can be formed as follows:

(a) choose an initial α.

(b) Find the $r_i^*(\alpha)$ $(i = n-1,\ n-2,\ \ldots\ 0)$ recursively, via numerical integration of expressions like (26) (where $r^*_{n-3}(\alpha)$ is defined by the condition $\Delta_{n-2}(r^*_{n-3}) = 0$).

(c) Compute

$$\mathfrak{L}_0^0(\alpha) = -e_1 + \int_0^\infty [(\beta_0-\alpha)\, r_0 + (\Delta_1(r_0))^+]\, f(r_0)\, dr_0.$$

(d) If $|\mathfrak{L}_0^0(\alpha)| < \epsilon$, for sufficiently small ϵ, say $L_{\max} = \alpha^* = \alpha$: otherwise repeat the computational cycle starting with a new α.

The following properties of $\mathfrak{L}_0^0(\alpha)$ can be used to generate an α-sequence which converges to α^*.

1. $\mathfrak{L}_0^0(\alpha)$ is monotone decreasing, since $\mathfrak{L}_0(\alpha)$ has this property for a fixed policy (see Eq. (19)); and if $\mathfrak{L}_0^0(\alpha_2) \geqslant \mathfrak{L}_0^0(\alpha_1)$ for $\alpha_2 > \alpha_1$, then the policy used to achieve $\mathfrak{L}_0^0(\alpha_2)$ could be used to achieve an $\mathfrak{L}_0(\alpha_1) > \mathfrak{L}_0^0(\alpha_1)$ — a contradiction.

2. When $\rho = 0$, all r_i^* are zero or infinite: replacement always occurs on arrival at a critical state i^*. Use of that policy will achieve the same average reward for durations having any value of ρ. Thus, a useful bound on $\alpha^*(\rho)$ is $\alpha^*(0) \leqslant \alpha^*(\rho)$; $0 \leqslant \rho \leqslant 1$.

3. When $\rho = 1$, future r_i are completely predictable ($Var(r_i|r_{i-1}) = 0$ in (A-7)), so $\alpha^*(1) \geqslant \alpha^*(\rho)$. In this case there is essentially a single random variable r_0, and the r_i^* can be calculated without the need for numerical integration of Bessel functions.

4. NUMERICAL EXAMPLE

Table I lists parameter values for a replacement problem which fall under the assumptions of Section 2.

TABLE 1 — *Numerical Example Parameters*

i	0	1	2	3	4	5
β_i	5	4	3	2	1	
η_i	1	0.9	0.8	0.7	0.6	
$E[p_i]$		2	2.2	2.4	2.6	2.8
$E[d_i]$		1	1.1	1.2	1.3	1.4

CASE 1 ($\rho = 0$)

Since future durations are independent of past ones, the optimal policy replaces when a critical state i^* is reached. The general optimal reward expression

$$\alpha^*(\rho) = \frac{E\left[\sum_0^{N-1} \beta_i r_i - p_N\right]}{E\left[\sum_0^{N-1} r_i + d_N\right]},$$

becomes, in this case

$$\alpha^*(0) = \max_j \left[\frac{\sum_0^{j-1} \beta_i \eta_i - E[p_i]}{\sum_0^{j-1} \eta_i + E[d_i]}\right] = \max_j A(j)$$

Direct evaluation shows

j	1	2	3	4	5
$A(j)$	1.5	2.13	**2.205**	2.085	1.89

with $j^* = 3$ and $\alpha^*(0) = 2.205$.

CASE 2 ($\rho = 1$)

Since $r_i = r_0\,\eta_i/\eta_0$ in this case, the optimal rule specifies a replacement state $j(r_0)$ as a function for r_0.

For any such policy

$$\mathfrak{L}_0(\alpha, j(r_0)) = E_{r_0}\left[-p_j - \alpha d_j + \frac{r_0}{\eta_0}\sum_0^{j-1}\eta_i(\beta_i - \alpha)\right].$$

This expectation will be maximized if $j(r_0)$ maximizes the bracketed term for each r_0. Making the necessary comparisons for a sequence of α-values leads to the policy

$$\begin{aligned} j^* &= 1, \text{ if } r_0 < 0.2698 \\ &= 2, \text{ if } 0.2698 \leqslant r_0 < 0.7083 \\ &= 3, \text{ if } 0.7083 \leqslant r_0. \end{aligned}$$

for which $|\mathfrak{L}_0| < 0.003$ and $\alpha^*(1) = 2.25$.

CASE 3 $\left(\rho = \frac{1}{2}\right)$

We know that $2.205 < \alpha^*\left(\frac{1}{2}\right) < 2.25$. A pilot calculation along the lines indicated in the previous section shows that $r_0^\bullet\left(\frac{1}{2}\right) = 0$, $r_j^\bullet\left(\frac{1}{2}\right) = \infty$ for $j \geqslant 2$, and

$$r_1^\bullet = \frac{9(\alpha^* - 2)}{8(3 - \alpha^*)},$$

where α^* is chosen to make the following $\mathfrak{L}_0(\alpha)$ vanish.

$$\mathfrak{L}_0(\alpha) = 6.4 - 3\alpha + \int_0^\infty \int_{r_1^\bullet}^\infty \left[\left(1 - \frac{\alpha}{2}\right) + \frac{4}{9}(3 - \alpha)r_1\right] \frac{e^{-\left(2r_0 + \frac{r_1}{0.45}\right)}}{0.45} I_0(2.981\sqrt{r_0 r_1})\,dr_1 dr_0$$

The known bounds on the optimal reward $\alpha^*\left(\frac{1}{2}\right)$ imply that the optimal threshold $r_1^\bullet$ is bounded, too: $0.290 < r_1^\bullet\left(\frac{1}{2}\right) < 0.375$.

Similar study of other values of the correlation parameter ρ lead to the optimal policy pattern described in Table II. One might say that as ρ increases, the past observations are more informative, the optimal policy makes finer distinctions, and the optimal reward increases.

5. CONCLUSIONS

A multivariate exponential distribution has been used to describe successive stages of deterioration. Optimal replacement strategies have been found for the class of decision rules which can continuously observe the deterioration state, and which may make replacements only

TABLE 2 — *Optimal Policy Structure*

Replacement State	Correlation Parameter ρ					
		0	1/4	1/2	3/4	1
	1				$r_0 < r_0^{\bullet}(3/4)$	$r_0 < r_0^{\bullet}(1)$
	2			$r_1 < r_1^{\bullet}(1/2)$	$r_1 < r_1^{\bullet}(3/4)$	$r_0 < \frac{\eta_0}{\eta_1} r_1^{\bullet}(1)$
	3	always	always	$r_1 \geqslant r_1^{\bullet}(1/2)$	$r_1 \geqslant r_1^{\bullet}(3/4)$	$r_0 \geqslant \frac{\eta_0}{\eta_1} r_1^{\bullet}(1)$

at the times of state transitions. Similar results have been found for the other reward rates shown in Figure 2 (linear; and constant after an initial set-up interval for readjustment to the new state) [5].

The optimal replacement policy derived in Section 2 makes use of observations which allow estimation of the current *rate* of deterioration for the correlated stages of deterioration. The numerical example demonstrates how the optimal policy and reward are related to the amount of correlation between the durations in successive deterioration states. For the model used here, the optimal policy for $\rho = 0$ will achieve the same reward (less than optimal) for any ρ. Depending on the application, the suboptimal approach may be satisfactory. The additional reward achievable by the actual optimal policy is bounded by the easily computed optimal reward for $\rho = 1$. However, it is possible that the small percentage improvement achievable for the $\rho = 1/2$ case in the example could represent a significant gain in a particular application.

The ordering of state dependent rewards, mean durations, etc. assumed here are physically reasonable, and lead to nice ordering of the decision regions. However, other β_i, η_i, p_i, d_i orderings might be more appropriate in other situations. The model introduced here for dependent stage durations could be used in those cases, together with dynamic programming optimization, although the solutions may not have comparably neat structures.

We anticipate that the optimization approach and policy structure described here will also be applicable to replacement problems having similar deterioration models. One easy extension would be to change the correlation structure in (A-3) from $\rho^{|i-j|}$ to something else, e.g., $\rho_1^{|i-j|} + \rho_2^{|i-j|}$. Other changes could permit the r_i to have non-exponential distributions, as long as similar total-positivity properties exist to permit analogous simplifications in the dynamic programming arguments.

Some of these other r_i distributions are being studied now in the hope of finding similar models which exhibit large percentage differences between the optimal rewards as $\rho_{r_i r_{i+1}}$ changes from zero to one. (Other choices of the numerical values in Table I have not revealed any such cases for the current model).

One reasonable generalization would allow transitions from state i to any state $j > i$. This would not change the form of the solution in the case of constant replacement penalties. How-

ever, the possibility of these additional transitions does ruin the structure when replacement penalties increase with the deterioration state. (The $\delta_{n-2}(s) > \delta_{n-1}(s)$ argument is no longer valid.)

REFERENCES

[1] Barlow, R.E. and F. Proschan, "Mathematical Theory of Reliability," John Wiley and Sons (1965).

[2] Barlow, R.E. and F. Proschan, "Statistical Theory of Reliability and Life Testing," Holt, Rinehart, and Winston (1975).

[3] Griffith, R.C., "Infinitely Divisible Multivariate Gamma Distributions," Sankhya, Series A,*32*, 393-404 (1970).

[4] Gumbel, E.J., "Bivariate Exponential Distributions," Journal of the American Statistical Association, *55*, 678-707 (1960).

[5] Hsu, C-L., L. Shaw and S.G. Tyan, "Reliability Applications of Multivariate Exponential Distributions," Technical Report, Poly-EE-77-036, Polytechnic Institute of New York (1977).

[6] Kao, E.P., "Optimal Replacement Rules when Changes of States are Semi-Markovian," Operations Research, *21*, 1231-1249 (1973).

[7] Karlin, S., "Total Positivity," Stanford University Press (1968).

[8] Kibble, W.F., "A Two-Variate Gamma Type Distribution," Sankhya, *5*, 137-150 (1941).

[9] Luss, H., "Maintenance Policies when Deterioration Can Be Observed by Inspections," Operations Research, *24*, 359-366 (1976).

[10] Marshall, A.W. and J. Olkin, "A Multivariate Exponential Distribution," Journal of the American Statistical Association, *22*, 30-44 (1967).

[11] Ross, S., "Applied Probability Models with Optimization Applications," Holden-Day (1970).

APPENDIX

Dependence Relationships Among Multivariate Exponential Variables

Many multivariate distributions have been described and applied to reliability problems [4,8,10]. In each case the marginal univariate distributions are of the negative exponential form. Properties of the distribution used here are most easily derived by exploiting its relationship to multivariate normal distributions [3.5].

The multivariate exponential variables $r_1, r_2 \ldots, r_n$ can be viewed as sums of squares:

(A-1) $$r_i = w_i^2 + z_i^2,$$

where $\underline{w}$ and $\underline{z}$ are independent, zero mean, identically distributed normal vectors, each with covariance matrix Γ. It follows that the r_i have exponential marginal distributions with

(A-2) $$E[r_i] = 2\gamma_{ii}$$

$$\rho_{r_i r_j} = [\rho_{w_i w_j}]^2.$$

We specialize to the case where the underlying normal sequences $\{w_i\}$ and $\{z_i\}$ are Markovian

(A-3) $$\gamma_{ij} = \sqrt{\gamma_{ii}\,\gamma_{jj}}\,\rho^{|i-j|}$$

and find that $\{r_i\}$ is also Markov with the joint density

$$(A\text{-}4)\qquad f(r_0,r_1,r_2,\ \ldots\ ,r_{n-1}) = \left[(1-\rho)^{n-1}\prod_{i=0}^{n-1}\eta_i\right]^{-1} \cdot \prod_{i=0}^{n-2} I_0\left[\frac{2}{1-\rho}\sqrt{\frac{\rho}{\eta_i\eta_{i+1}}}\sqrt{r_i r_{i+1}}\right] \cdot \exp\left[-\frac{1}{1-\rho}\left(\frac{r_0}{\eta_0}+\frac{r_{n-1}}{\eta_{n-1}}+\sum_{i-1}^{n-2}\frac{r_i(1+\rho)}{\eta_i}\right)\right];\ n>2,$$

Equation (A-4) uses the modified Bessel function $I_0(\ )$ and the notations $E[r_i]=\eta_i$ and $\rho_{r_i,r_{i+1}}=\rho$. (When $n=2$, the summation in exp ( ) vanishes.)

The conditional density is easily shown to satisfy the Markov property and [5]

$$(A\text{-}5)\qquad f(r_i|r_{i-1}) = [\eta_i(i-\rho)]^{-1}\exp\left[-\frac{1}{(1-\rho)}\left(\frac{r_i}{\eta_i}+\frac{\rho r_{i-1}}{\eta_{i-1}}\right)\right] \cdot I_0\left(\frac{2}{1-\rho}\sqrt{\frac{\rho r_i r_{i-1}}{\eta_i\eta_{i-1}}}\right)$$

with

$$(A\text{-}6)\qquad E[r_i|r_{i-1}] = \eta_i + (r_{i-1}-\eta_{i-1})\rho\ \eta_i/\eta_{i-1}.$$

$$(A\text{-}7)\qquad \mathrm{Var}[r_i|r_{i-1}] = \eta_i^2[(1-\rho)^2 + 2\rho(1-\rho)r_{i-1}/\eta_{i-1}].$$

These conditional moments shows, e.g., that the conditional mean of r_i exceeds its mean in proportion to the amount by which r_{i-1} exceeds its mean, and that conditional mean is a linearly increasing function of r_{i-1}.

The dynamic programming arguments used here required calculations of conditional expectations based on (A-5). As is often the case [2], the total positivity properties of $f(r_i|r_{i-1})$ are very useful for determining structural properties of the optimal policy.

It is straightforward to show that both $f(r_i,\ r_{i-1})$ and $f(r_i|r_{i-1})$ are totally positive of all orders (TP_∞), [5,7]. This means, for $f(r_i,\ r_{i-1})$, that the following determinants are nonnegative for any N and any $\alpha_1 < \alpha_2 \ldots < \alpha_N$; $\beta_1 < \beta_2 \ldots < \beta_N$.

$$\begin{vmatrix} f(\alpha_1,\beta_1) & f(\alpha_1\cdot\beta_2) & \ldots & f(\alpha_1,\beta_N) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ f(\alpha_N,\beta_1) & \ldots\ldots\ldots\ldots & & f(\alpha_N,\beta_N) \end{vmatrix} \geqslant 0.$$

THEOREM: if $h(y)$ is continuous and convex, and satisfies the bounds

(i) $h(0) \leqslant 0$

(ii) $|h(y)| \leqslant a + b\,y^{2m}$; $a>0$, $b>0$, $y>0$, m = positive integer. $g(x) = \int h(y)\,f(y|x)\,dy$, and $f(y|x)$ is TP_∞, then $g(x)$ is continuous, convex, bounded in the sense

$$|g(x)| \leqslant a' + b'\,x^{2m};\ a'>0,\ b'>0,\ x>0;$$

and belongs to one of the three following categories:

(I) $g(x) \geqslant 0$ for all $x \geqslant 0$,

(II) $g(x) < 0$ for all $x \geqslant 0$ except with a possible zero at $x = 0$,

(III) there exists a unique x^*, $0 < x^* < \infty$, such that $g(x) > 0$ for all $x > x^*$; and $g(x) < 0$ for $x < x^*$ except for a possible zero at $x = 0$.

This theorem is used to define optimal decision regions according to the sign of a function like $g(x)$, with x^* corresponding to a decision threshold.

STATISTICAL ANALYSIS OF A CONVENTIONAL FUZE TIMER

Edgar A. Cohen, Jr.

Naval Surface Weapons Center
White Oak
Silver Spring, Maryland

ABSTRACT

In this paper, a statistical analytic model for evaluation of the performance of a standard electric bomb fuze timer is presented. The model is based on what is called a selective design assembly, where one item, namely, a resistor, is used to time the circuit. In such an assembly, the remaining components are chosen a priori from predetermined distributions. Based on the analysis, a general numerical integration scheme is utilized for assessing performance of the timer. The results of a computer simulation are also given. In the last section of the paper, a theory for evaluation of the yield of two or more timers designed to operate in sequence is derived. To appraise such a scheme, a numerical quadrature routine is developed.

1. INTRODUCTION AND PHYSICAL DESCRIPTION

In this paper, we shall be concerned with the statistical analysis of the bomb fuze timer shown in Figure 1. As is common in practice, a standard, or precision, resistor is used to time the circuit after the rest of the components have been assembled in a random fashion. Then, to meet certain timing requirements to be discussed later, a resistor is selected and introduced into the circuit. A number of tests must afterwards be performed in sequence to check the performance of the product under differing environmental conditions. Such environmental influences are, for example, temperature effects, effect of packaging, resistor incrementation (to be discussed), and effect of vibration and moisture uptake. In addition, one might have several timers which operate sequentially, all fed from the same energy storage capacitor C1 of Figure 1. This paper is devoted to an analysis of such a timer in what is called the ambient temperature range, whose limits are 70°F and 80°F, respectively. We will also indicate the procedure for treating analytically the assessment of performance of combinations of several timers. The author has been involved in a Monte Carlo study for the Navy of such timers. Previous work has involved reliability studies of an entire fuze assembly using these timers [2].

2. RESISTOR SELECTION PROCESS

The timer indicated in Figure 1 works once the potential difference across the two capacitors $C2$ and $C3$ is sufficient to fire the cold cathode diode tube *VT*. Capacitors $C1$ and $C3$ initially have the same potential across them. As time progresses, $C1$ discharges through resistor RES into $C2$, while $C3$ serves as a reference capacitor. Thus, the voltage across $C2$ builds up

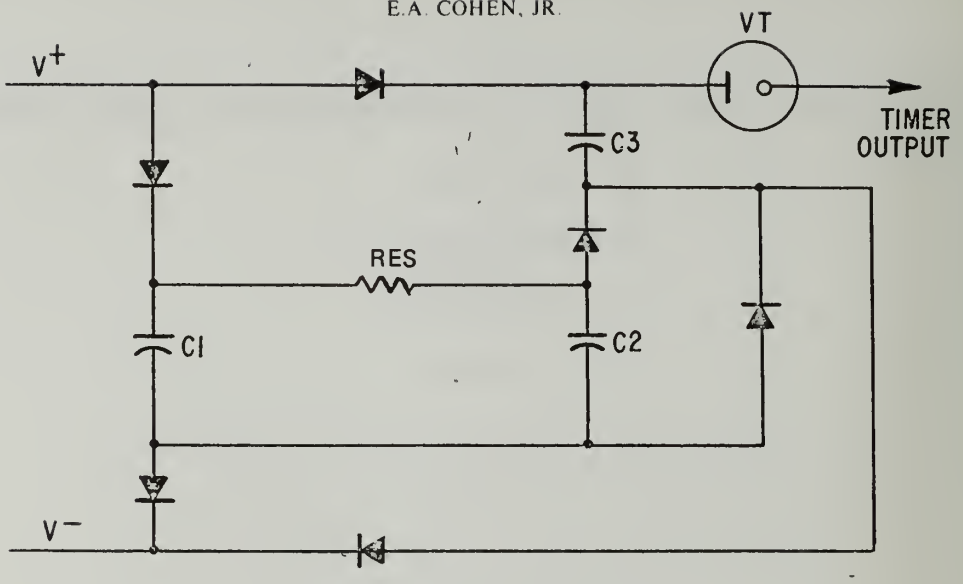


FIGURE 1. Fuze timer configuration

until the potential across tubes $C2$ and $C3$ is adequate to fire tube VT. The relationship between firing time and the values of the circuit components can be derived from a simple first order differential equation and is given by

$$t = \frac{RC_1C_2}{C_1 + C_2} \ln\left[\frac{VC_1}{VC_1 - (V_T - V)(C_1 + C_2)}\right], \tag{2.1}$$

where

C_1 = capacitance of capacitor $C1$

C_2 = capacitance of capacitor $C2$

V = supply voltage (potential across $C3$ and potential initially across $C1$)

V_T = firing voltage of cold cathode diode tube VT

R = resistance of resistor RES.

To illustrate the pertinent features of the process, write (2.1), for brevity, in the form

$$t = \mathrm{RF}(C_1, C_2, V, V_T). \tag{2.2}$$

Note that (2.2) is linear and homogeneous in R, so that R can be used as a scaling parameter. This is precisely how it is used when the timer is first assembled.

In practice, the resistors are supplied in large numbers by the manufacturer, after which they are tested and sorted by the user into a large number of bins. The resistors in each bin have resistances, at a standard temperature, which fall into a certain interval. These intervals are arranged to have the same "percent width", to be described in more detail below. The timer is to be designed to fire at a nominal time t_N. Since capacitors $C1$ and $C2$ are chosen at random from a lot, their capacitances C_1 and C_2 may be treated as random variables. Likewise, tube firing voltage V_T may also be considered as a random variable. In general, we shall also consider the supply voltage V to be random.

Let us agree to denote by R_0 that value of R obtained from relation (2.1) when $t = t_N$ and C_1, C_2, V, and V_T are given their expected values at some standard temperature, e.g., 75°F. For convenience, R_0 may be used as a reference resistance, and the bin to which reference resistor RES_0, of resistance R_0, belongs could be called the reference resistor bin. The interval corresponding to this bin is to contain all resistances which fall between $R_0(1-\epsilon)$ and $R_0(1+\epsilon)$, where ϵ is a preassigned small positive number. Our second bin will contain all resistors whose resistances fall between $R_0(1+\epsilon)$ and $R_0(1+\epsilon)^2/(1-\epsilon)$, and the third bin those resistors whose resistances lie between $R_0(1-\epsilon)^2/(1+\epsilon)$ and $R_0(1-\epsilon)$. In general, our intervals are to be so constructed that the ratio of right endpoint to left endpoint is always $(1+\epsilon)/(1-\epsilon)$, which, to first order accuracy, is just $1+2\epsilon$. Alternatively, one may divide the difference of the two endpoints by its midpoint to obtain precisely 2ϵ. We shall, therefore, say that each such interval has "percent width" 2ϵ. In setting up the interval division scheme, a percent increment ϵ_1 is chosen a priori, and then $\epsilon = \epsilon_1/100$. This ϵ_1 is typically of the order of 1/2 to 1%. Figure 2 is a diagram of this scheme.

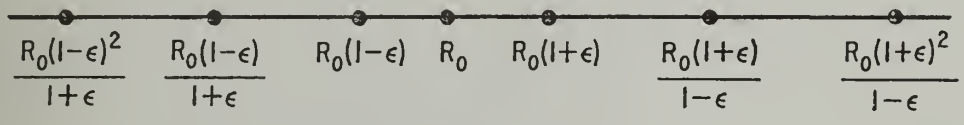


FIGURE 2. Resistance interval setup

Once again referring to our circuit configuration, where C_1, C_2, V, and V_T are random variables, let us define

$$t_0 = R_0 F(C_1, C_2, V, V_T). \tag{2.3}$$

Then, to achieve the nominal time t_N, we define our nominal resistance to be

$$R_N = R_0 t_N / t_0. \tag{2.4}$$

Note that, since t_0 is a random variable (being a function of the random variables C_1, C_2, V, and V_T), R_N is also a random variable. A technician may use relation (2.4) to determine R_N. Then he picks a resistor RES_p at random from the bin to which resistor RES_N belongs and integrates such resistor, of resistance R_p, into the circuit. This process is called, in fuze technology parlance, "resistor incrementation." Note that R_p is a random variable which is statistically dependent on R_N inasmuch as R_p and R_N must lie in the same interval. However, once attention is restricted to a given interval of the scheme, it is clear that the value of R_N in no way influences the value of R_p, since one is free to select any resistor in the bin to which the nominal resistor belongs. We shall reemphasize this fact in Section 3. For simplicity we index the intervals by i, letting their left and right endpoints be r_i and r_{i+1}, respectively. To achieve compatibility, the bins should initially be formed and kept at some standard temperature, and the timer should be assembled at that same temperature. In practice, this will, in all likelihood, not be the case, but one may compensate for this defect by studying the sensitivity of the timer to changes in bin interval width. For example, if by doubling the interval width, the overall change in performance is insignificant, it may be safely assumed that such a discrepancy was unimportant (provided the distributions due to ambient temperature variations are of small variance).

3. PROBABILITY INTERVALS AT THE STANDARD TEMPERATURE

The problem of determining the probability of operation of the timer within two given times, say t_1 and t_2, when there is no effect other than resistor selection is not difficult. (We also ignore, in this section, the effect of tube firing voltage variation from one firing to the

next. This phenomenon will be discussed in some detail in Section 4.) The reason is that the time is linear and homogeneous in resistance R. In fact, the bins have been designed to take advantage of this feature, and we shall show that the probability interval is independent of the bin in which resistor RES_N falls.

First of all, let $t_{\min}^{(i)}$ and $t_{\max}^{(i)}$ be the minimum and maximum times, respectively, obtainable when the nominal resistance R_N and the picked resistance R_p come from a given bin i. Also, let $F_{\min}^{(i)}$ and $F_{\max}^{(i)}$ be the smallest and largest values of F, respectively, given only t_N and knowing that R_N comes from that bin. It follows that

$$t_{\min}^{(i)} = r_i F_{\min}^{(i)} = r_i t_N / r_{i+1} \tag{3.1}$$

and

$$t_{\max}^{(i)} = r_{i+1} F_{\max}^{(i)} = r_{i+1} t_N / r_i. \tag{3.2}$$

Therefore, given that R_N and R_p lie in interval i,

$$r_i t_N / r_{i+1} \leqslant t \leqslant r_{i+1} t_N / r_i. \tag{3.3}$$

Since $$r_i / r_{i+1} = (1 - \epsilon)/(1 + \epsilon),$$

$$(1 - \epsilon)/(1 + \epsilon) \leqslant t/t_N \leqslant (1 + \epsilon)/(1 - \epsilon), \tag{3.4}$$

independent of bin interval. In other words, (3.4) is true with probability 1.

Generally, suppose that one is interested in the probability that firing time falls between two prescribed limits about the nominal time. Consider once more a given bin i. Let us denote by $R_N^{(i)}$ and $R_p^{(i)}$ random variables derived from R_N and R_p respectively under the condition that R_N and, therefore, R_p must lie in interval i. From our discussion in section 2, it is clear that these new random variables must be independent. Let t_1 and t_2 be the lower and upper limits, respectively, on firing time. For any given value of the random variable $R_N^{(i)}$, one can determine limits on the random variable $R_p^{(i)}$ so that the firing time lies between t_1 and t_2. Since, by definition, $t_N = R_N^{(i)} F$, it follows that $R_p^{(i)}$ cannot be less than

$$t_1 / F = t_1 R_N^{(i)} / t_N. \tag{3.5}$$

Similarly, $R_p^{(i)}$ cannot exceed

$$t_2 R_N^{(i)} / t_N. \tag{3.6}$$

One must, of course, realize that (3.5) may be smaller than r_i and (3.6) larger than r_{i+1} for values of $R_N^{(i)}$ close to r_i and r_{i+1}, respectively.

If we let $g(R_N)$ be the density function of the random variable R_N defined by (2.3) and (2.4), whose range is a function of the domain of C_1, C_2, V, and V_T, then the induced random variable $R_N^{(i)}$ has conditional density

$$g^{(i)}(R_N^{(i)}) = g(R_N) / P(r_i \leqslant R_N \leqslant r_{i+1}) = g(R_N) / \int_{r_i}^{r_{i+1}} g(R_N)\, dR_N. \tag{3.7}$$

The range of $R_N^{(i)}$ is restricted to the interval $[r_i, r_{i+1}]$. Using the mean value theorem of integral calculus, (3.7) becomes

$$g^{(i)}(R_N^{(i)}) = g(R_N) / g(\xi)(r_{i+1} - r_i), \quad r_i \leqslant \xi \leqslant r_{i+1}. \tag{3.8}$$

If $r_{i+1} - r_i$ is sufficiently small, one sees that

$$g^{(i)}(R_N^{(i)}) \cong 1/(r_{i+1} - r_i). \tag{3.9}$$

Similarly, let $f^{(i)}(R_p^{(i)})$ be the density function for picked resistance $R_p^{(i)}$, whose range is likewise restricted to $[r_i, r_{i+1}]$. Then, with the knowledge that $R_N^{(i)}$ and $R_p^{(i)}$ are independent random variables, and, letting $P_i(t_1 \leqslant t \leqslant t_2)$ be the probability that firing time falls between t_1 and t_2 (given that R_N and R_p come from interval i),

$$P_i(t_1 \leqslant t \leqslant t_2) = \int_{r_i}^{r_{i+1}} \int_{t_1 R_N^{(i)}/t_N}^{t_2 R_N^{(i)}/t_N} g^{(i)}(R_N^{(i)}) f^{(i)}(R_p^{(i)})\, dR_p^{(i)}\, dR_N^{(i)}. \tag{3.10}$$

We take the liberty of defining $f^{(i)}(R_p^{(i)}) = 0$ in (3.10) whenever $R_p^{(i)} \notin [r_i, r_{i+1}]$. This is done purely for the sake of convenience of notation even though the range of $R_p^{(i)}$ is $[r_i, r_{i+1}]$.

The probability that the time falls between t_1 and t_2 is expressed by

$$P(t_1 \leqslant t \leqslant t_2) = \sum_{i=-\infty}^{\infty} p_i P_i(t_1 \leqslant t \leqslant t_2), \tag{3.11}$$

where p_i is the probability of choosing bin i.

As we have previously indicated, if $r_{i+1} - r_i$ is sufficiently small, we can assume, for all practical purposes, that $R_N^{(i)}$ is a uniformly distributed random variable. The picked resistance $R_p^{(i)}$ should also be a uniformly distributed random variable if all resistors in bin i are equally likely to appear. In other words, let us assume that

$$g^{(i)}(R_N^{(i)}) = f^{(i)}(R_p^{(i)}) = 1/(r_{i+1} - r_i). \tag{3.12}$$

Suppose then that one asks for the probability that $t_1 = t_N(1 - \delta) \leqslant t \leqslant t_N(1 + \delta) = t_2$ for a given small δ. We proceed to derive closed form expressions for this probability. Three cases naturally arise, the first of which is shown in Figure 3 below. For brevity, we shall drop the superscript i in this figure and the two following figures. In this diagram, the interior of the quadrilateral formed by the lines $R_N = r_i$, $R_N = r_{i+1}$, $R_p = t_1 R_N/t_N$, and $R_p = t_2 R_N/t_N$ is the region of integration. Note that, in the two hatched regions, $f^{(i)}(R_p^{(i)}) = 0$, since then either $R_p < r_i$ or $R_p > r_{i+1}$. After a small computation, one sees that the inequality $r_N^{(0)} < r_N^{(1)}$ is equivalent to

$$0 < \delta < \epsilon. \tag{3.13}$$

We also note that, using (3.12), (3.10) represents the normalized area of the interior of the hexagon shown in Figure 3, bounded by the lines $R_N = r_i$, $R_N = r_{i+1}$, $R_p = t_1 R_N/t_N$, $R_p = t_2 R_N/t_N$, $R_p = r_i$, and $R_p = r_{i+1}$. Therefore,

$$\begin{aligned} &P_i(t_N(1 - \delta) \leqslant t \leqslant t_N(1 + \delta)) \\ &= \frac{1}{(r_{i+1} - r_i)^2} \left[\int_{r_i}^{r_i/(1-\delta)} \int_{r_i}^{(1+\delta)R_N^{(i)}} dR^{(i)}\, dR_N^{(i)} \right. \\ &+ \int_{r_i/(1-\delta)}^{r_{i+1}/(1+\delta)} \int_{(1-\delta)R_N^{(i)}}^{(1+\delta)R_N^{(i)}} dR^{(i)}\, dR_N^{(i)} \\ &\left. + \int_{r_{i+1}/(1+\delta)}^{r_{i+1}} \int_{(1-\delta)R_N^{(i)}}^{r_{i+1}} dR^{(i)}\, dR_N^{(i)} \right] \\ &= \frac{\delta}{8\epsilon^2} \left[\frac{(1 + \epsilon)^2(2 + \delta)}{1 + \delta} - \frac{(1 - \epsilon)^2(2 - \delta)}{1 - \delta} \right], \quad 0 < \delta < \epsilon. \end{aligned} \tag{3.14}$$

It follows that P_i is independent of i. From (3.11),

$$P(t_1 \leqslant t \leqslant t_2) = P_i(t_1 \leqslant t \leqslant t_2). \tag{3.15}$$

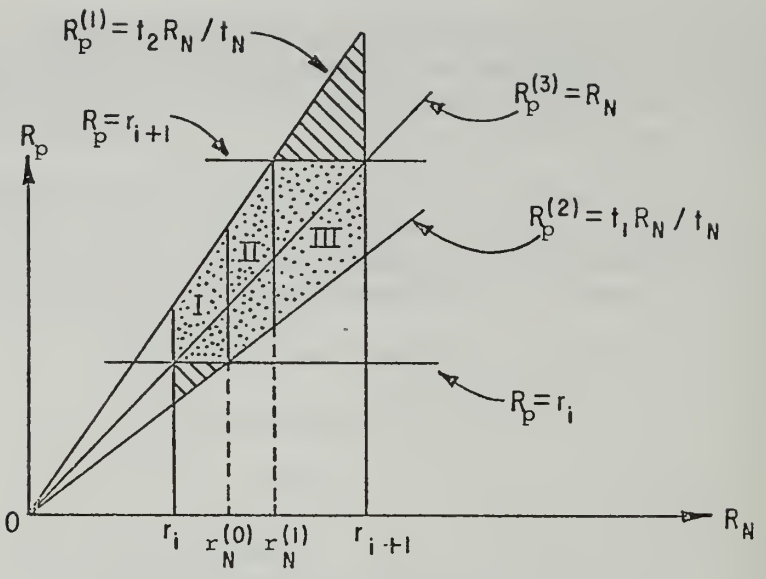


FIGURE 3. Picked resistance versus nominal resistance (Region 1)

The second case occurs when $r_i \leqslant r_N^{(1)} \leqslant r_N^{(0)} \leqslant r_{i+1}$. This situation is indicated in Figure 4. One can also show that $r_N^{(0)} = r_{i+1}$ when $\delta = 2\epsilon/(1+\epsilon)$ and that $r_N^{(1)} = r_i$ when $\delta = 2\epsilon/(1-\epsilon)$. Therefore, the situation illustrated in Figure 4 occurs when $\epsilon \leqslant \delta \leqslant 2\epsilon/(1+\epsilon)$. A third case will occur when $2\epsilon/(1+\epsilon) < \delta < 2\epsilon/(1-\epsilon)$, as illustrated in Figure 5, where the dotted region is now a pentagon. For $\delta \geqslant 2\epsilon/(1-\epsilon)$, the dotted region becomes the interior of a rectangle completely enclosed in the sector, so that the probability becomes unity. In the third case, one sees that $r_i \leqslant r_N^{(1)} \leqslant r_{i+1} \leqslant r_N^{(0)}$. When one integrates over the interior of the quadrilateral outlined in Figure 4, one again obtains the closed form given in (3.14). Therefore, (3.14) is valid whenever $0 \leqslant \delta \leqslant 2\epsilon/(1+\epsilon)$. The case illustrated in Figure 5 is different. When we integrate over the interior of the pentagon, which is that portion of the region of integration for which the integrand of (3.10) is nonzero, we find that

$$P_i(t_N(1-\delta) \leqslant t \leqslant t_N(1+\delta)) = \frac{4\epsilon^2 + 4\epsilon(1+\epsilon)\delta - (1-\epsilon)^2\delta^2}{8\epsilon^2(1+\delta)}, \tag{3.16}$$

$$\frac{2\epsilon}{1+\epsilon} \leqslant \delta \leqslant \frac{2\epsilon}{1-\epsilon}.$$

One easily shows that (3.16) becomes unity when $\delta = 2\epsilon/(1-\epsilon)$ is substituted.

4. PROBABILITY INTERVALS AT AMBIENT TEMPERATURE BEFORE POTTING

The analysis of the timer when temperature and cold cathode diode firing voltage variations are considered is different from that of the previous section, since all components except for the resistor enter the time nonlinearly. It would then be necessary, at least in principle, to take into consideration the probabilities p_i of picking the bins as well as the probabilities for picked resistance once a bin has been selected. However, if the variations due to these effects are relatively small, one should again see probabilities essentially independent of the bin selected. Furthermore, in a situation like this wherein certain distributions are quite tight, i.e., are of small variance, some simplifying assumptions can be made. We shall get to these presently. Again, as before, we assume that the bin intervals are so small that we may reasonably suppose that (3.12) is true. Note also that (2.3) and (2.4) express R_N in terms of t_N, C_1, C_2, V, and V_T. Assume now that C_1, C_2, V, and V_T are independent, normally distributed random variables. Suppose, as is common in practice when coefficients of variation are

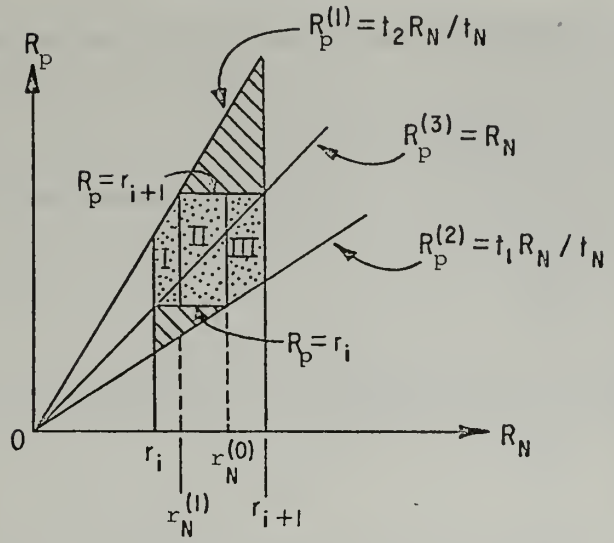


FIGURE 4. Picked resistance versus nominal resistance (Region 2)

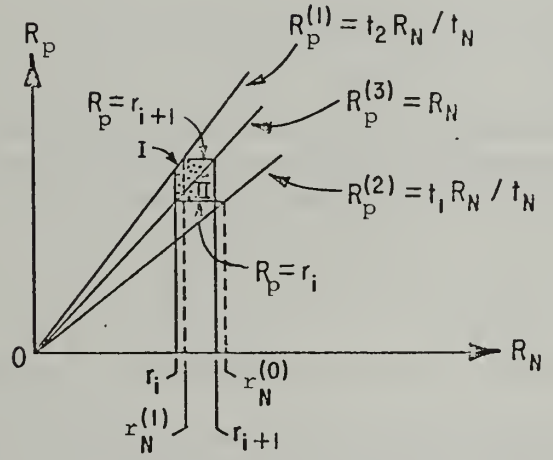


FIGURE 5. Picked resistance versus nominal resistance (Region 3)

small [4, pp. 246-251], that R_N is linearized about the expected values of capacitances C_1 and C_2, supply voltage V, and tube breakdown voltage V_T. Now a linear function of independent, normally distributed random variables is again a normally distributed random variable, and, from (2.3) and (2.4), it follows [4, pg. 118] that

$$E(R_N) \cong \frac{t_N(C_{1,E} + C_{2,E})}{C_{1,E}C_{2,E}} \left\{ \ell n \left[\frac{V_E C_{1,E}}{V_E C_{1,E} - (V_{T,E} - V_E)(C_{1,E} + C_{2,E})} \right] \right\}^{-1} \tag{4.1}$$

and

$$\begin{aligned} \operatorname{var}(R_N) \cong & \left(\frac{\partial R_N}{\partial C_1} \right)_E^2 \operatorname{var} C_1 + \left(\frac{\partial R_N}{\partial C_2} \right)_E^2 \operatorname{var} C_2 \\ & + \left(\frac{\partial R_N}{\partial V} \right)_E^2 \operatorname{var} V + \left(\frac{\partial R_N}{\partial V_T} \right)_E^2 \operatorname{var} V_T. \end{aligned} \tag{4.2}$$

Here the subscript E indicates evaluation at expected values and var represents the variance operator. Now, clearly,

$$\frac{\partial R_N}{\partial C_i} = -\frac{t_N}{F^2}\frac{\partial F}{\partial C_i}, \quad i = 1, 2, \tag{4.3}$$

with similar expressions for $\partial R_N/\partial V$ and $\partial R_N/\partial V_T$. The relevant partial derivatives of F are given by

$$\begin{aligned}
\frac{\partial F}{\partial C_1} &= \frac{C_2^2}{C_1 + C_2}\left[\frac{1}{C_1 + C_2}X - \frac{V_T - V}{Y}\right] \\
\frac{\partial F}{\partial C_2} &= \frac{C_1}{C_1 + C_2}\left[\frac{C_1}{C_1 + C_2}X + \frac{C_2(V_T - V)}{Y}\right] \\
\frac{\partial F}{\partial V_T} &= \frac{C_1 C_2}{Y} \\
\frac{\partial F}{\partial V} &= \frac{C_1 C_2 V_T}{VY},
\end{aligned} \tag{4.4}$$

and

$$\begin{aligned}
X &= ln\left[\frac{VC_1}{VC_1 - (V_T - V)(C_1 + C_2)}\right], \\
Y &= VC_1 - (V_T - V)(C_1 + C_2).
\end{aligned}$$

Now p_i represents the probability of choosing bin i, and that is precisely the probability that the random variable R_N belongs to bin i. Furthermore, because we are now assuming that R_N is a linear function of the independent, normal random variables C_1, C_2, V, and V_T, R_N is likewise normal. Therefore, letting $\xi = E(R_N)$ and $\sigma^2 = \text{var}(R_N)$, one has

$$p_i = \frac{1}{\sigma\sqrt{2\pi}}\int_{r_i}^{r_{i+1}} e^{-\frac{1}{2}\left[\frac{r-\xi}{\sigma}\right]^2} dr = \frac{1}{\sqrt{2\pi}}\int_{v_1}^{v_2} e^{-\frac{1}{2}v^2} dv, \tag{4.5}$$

where $v_1 = (r_1 - \xi)/\sigma$ and $v_2 = (r_{i+1} - \xi)/\sigma$, so that p_i may be readily calculated from tables.

Supposing that the picked resistor and the other components are subject to a temperature change from the standard temperature, we must compute the effect of such a change, together with the resistor incrementation effect of Section 3, in order to obtain the probability of satisfying the specification. It will be assumed in our analysis that the ambient temperature is a uniformly distributed random variable whose range is given by $T_1 \leqslant T \leqslant T_2$. If $P(t_1 \leqslant t \leqslant t_2 | T)$ is the probability of meeting the time limits for a given temperature T, then, clearly,

$$P(t_1 \leqslant t \leqslant t_2) = \int_{T_1}^{T_2} P(t_1 \leqslant t \leqslant t_2 | T) p(T)\, dT = \frac{1}{T_2 - T_1}\int_{T_1}^{T_2} P(t_1 \leqslant t \leqslant t_2 | T)\, dT. \tag{4.6}$$

Let us give an example of the computation of the nominal resistor distribution. Supposing in (4.1) that $t_N = 2.6$ seconds, $C_{1,E} = .44\,\mu f$, $C_{2,E} = .15\,\mu f$, $V_E = 177$v., and $V_{T,E} = 235$v., one finds that $E(R_N) = 40.16$ megohms. Also, one finds from (4.3) and similar expressions, upon inserting expected values, that $\partial R_N/\partial C_1 = 8.65$, $\partial R_N/\partial C_2 = -293.12$, $\partial R_N/\partial V = -1.26$, and $\partial R_N/\partial V_T = -0.95$. Let us assume the following standard deviations: $\sigma(C_1) = 0.0073$, $\sigma(C_2) = 0.0025$, $\sigma(V) = 0.17$, and $\sigma(V_T^E) = 4.17$, where V_T^E is used to denote the expected breakdown voltage of a diode chosen from a lot. The expected values of the breakdown voltages of all the tubes are themselves assumed to follow a normal distribution with expected

value 235v. and with the above σ. In addition, each tube has a firing voltage which varies about its expected value. This new random variable, with expected value 0, we denote by ΔV_E, and it is assumed that ΔV_E is also normally distributed. The random variable V_T, which represents the firing voltage of a tube selected from a lot, is actually formed as a sum $V_T = V_T^E + \Delta V_E$, where we shall suppose that ΔV_E is independent of V_T^E. Also, tests performed by fuze specialists indicate that the random variables ΔV_E have the same distribution from one tube to the next. Assuming that $\sigma(\Delta V_E) = 0.24$, it follows that $\sigma(V_T) = 4.17$. Then, from (4.2), var $R_N = 16.3235$, or $\sigma(R_N) = 4.04$. Therefore, the coefficient of variation is 0.10, which is reasonably small.

We now develop a general method for evaluating the performance of the timer which is based on a linear theory. Hopefully, this theory will yield at least conservative estimates. Our formula is a generalization of that given in paragraph 3. First of all, from (2.3) and (2.4), it follows that

$$R_N = t_N / F(C_1, C_2, V, V_T^{(1)}), \tag{4.7}$$

where $V_T^{(1)} = V_T^E + \Delta V_E^{(1)}$. Therefore, solving (4.7) for V_T^E, where $F(C_1, C_2, V, V_T^{(1)})$ is given through (2.1) and (2.2), one finds that

$$V_T^E = VC_1/(C_1 + C_2) - (\Delta V_E^{(1)} - V) - VC_1 e^{-t_N/R_N C_{\text{eff.}}}/(C_1 + C_2). \tag{4.8}$$

Here $C_{\text{eff.}} \equiv 1/C_1 + 1/C_2$ is the effective series capacitance of C_1 and C_2, and $\Delta V_E^{(1)}$ denotes that variation in tube firing voltage from its expected value which is associated with determination of the nominal resistance R_N. For brevity, we let $g(C_1, C_2, V, R_N, \Delta V_E^{(1)})$ represent the right hand side of (4.8). There is, however, a second variation, which we shall denote by $\Delta V_E^{(2)}$, that occurs once a resistor has been selected from a bin and the timer actually operated. These two variations must be taken into account carefully when assessing timer performance. One may now make a 1-1 transformation from the space of $(C_1, C_2, V, \Delta V_E^{(1)}, \Delta V_E^{(2)}, V_T^E)$ to that of $(C_1, C_2, V, \Delta V_E^{(1)}, \Delta V_E^{(2)}, R_N)$ through the map

$$\begin{aligned} &C_1 = C_1,\ C_2 = C_2,\ V = V,\ \Delta V_E^{(1)} = \Delta V_E^{(1)},\ \Delta V_E^{(2)} = \Delta V_E^{(2)}, \\ &V_T^E = g(C_1, C_2, V, R_N, \Delta V_E^{(1)}), \end{aligned} \tag{4.9}$$

whose Jacobian is $\partial V_E/\partial R_N$. It follows [3, pp. 56-62] that the density function for the state $(C_1, C_2, V, \Delta V_E^{(1)}, \Delta V_E^{(2)}, R_N)$ is

$$\begin{aligned} f(C_1, C_2, V, \Delta V_E^{(1)}, \Delta V_E^{(2)}, R_N) = {} & p_1(C_1)p_2(C_2)p_3(V)p_4(\Delta V_E^{(1)}) \\ & \cdot p_5(\Delta V_E^{(2)})p_6[g(C_1, C_2, V, R_N, \Delta V_E^{(1)}] \\ & \cdot |\partial V_E/\partial R_N|, \end{aligned} \tag{4.10}$$

where $p_i(C_i)$ $(i = 1, 2)$ are the densities for C_i, p_3 is the density for V, p_4 the density for $\Delta V_E^{(1)}$, p_5 the density for $\Delta V_E^{(2)}$, and p_6 the density for V_T^E. These random variables are all assumed to be independent. In addition, $\Delta V_E^{(1)}$ and $\Delta V_E^{(2)}$ are identically distributed. Next account must be taken of the fact that, because of a change in temperature, the capacitances C_i will change in value. In fact, we assume that $C_i(T)$, where T denotes temperature, is of the form

$$C_i(T) = C_i(1 + K_i(T - T_E)/100), \tag{4.11}$$

where K_i represents a random percent change per degree from the expected temperature T_E. Thus $C_i(T)$ is a product convolution [3, pp. 56-62] of C_i and the second factor, which we denote by $\Delta CP_i(T)$ (representing a percentage change in C_i due to a temperature change from expected value T_E to T). We then form the joint density $h(C_1, C_2, \Delta CP_1(T), \Delta CP_2(T), V,$

$\Delta V_E^{(1)}$, $\Delta V_E^{(2)}$, R_N) from f and the densities for these percent changes. Afterwards, h is multiplied by $p(R_p(T))$, the convolution density of picked resistance at temperature, where

$$R_p(T) = R_p(1 + C(T - T_E)/100) \tag{4.12}$$

and C is a random percent change per degree. Finally, if we are interested in the conditional density for any given bin i, we must divide by p_i, the probability of choice of bin i. It is clear that, in order for the time output of the timer to fall between two chosen values t_1 and t_2, $R_p(T)$ must lie between

$$t_1/F(C_1(T),\ C_2(T),\ V,\ V_T^{(2)})$$

and

$$t_2/F(C_1(T),\ C_2(T),\ V,\ V_T^{(2)}),$$

where $V_T^{(2)} = V_T^E + \Delta V_E^{(2)}$ with V_T^E given by (4.8). Also, from (4.11),

$$C_i(T) = C_i \Delta CP_i(T). \tag{4.13}$$

Now let $X_T = (C_1,\ C_2,\ \Delta CP_1(T),\ \Delta CP_2(T),\ V,\ \Delta V_E^{(1)},\ \Delta V_E^{(2)})$. There follows the general multiple integration formula, which expresses the probability P_i that the time falls between t_1 and t_2 for bin i and conditioned on temperature T:

$$p_i P_i(t_i \leqslant t \leqslant t_2 | T) = \int_{r_i}^{r_{i+1}} \int_{X_T \in R^7(-\infty,\infty)} \int_{t_1/F}^{t_2/F} p(R_p(T))\, h(X_T, R_N)\, dR_p(T)\, dX_T\, dR_N, \tag{4.14}$$

where $R^7(-\infty, \infty)$ represents the seven-fold Cartesian product of the real line. Finally,

$$P(t_1 \leqslant t \leqslant t_2) = \frac{1}{T_2 - T_1} \sum_{-\infty}^{\infty} p_i \int_{T_1}^{T_2} P_i(t_1 \leqslant t \leqslant t_2 | T)\, dT, \tag{4.15}$$

given that the temperature distribution is uniform. This integration procedure could be accomplished on a digital computer through use of numerical Gaussian quadrature and Gauss-Hermite quadrature [5, pp. 130-132]. However, instead of using this general nonlinear approach, we find it convenient, in the present context, to linearize the products given by (4.11) and (4.12) and to make use of a linearized version of R_N given by

$$\begin{aligned} R_N = {} & R_N(C_{1,E},\ C_{2,E},\ V_E,\ V_{T,E}^{(1)}) + A_1(C_1 - C_{1,E}) + A_2(C_2 - C_{2,E}) \\ & + A_3(V - V_E) + A_4(V_T - V_{T,E}^{(1)}), \end{aligned} \tag{4.16}$$

where, of course,

$$A_1 = \frac{\partial R_N}{\partial C_1},\ A_2 = \frac{\partial R_N}{\partial C_2},\ A_3 = \frac{\partial R_N}{\partial V},\ A_4 = \frac{\partial R_N}{\partial V_T}$$

are evaluated at the expected values for the components and $V_{T,E}^{(1)}$ represents the expected value of random variable $V_T^{(1)}$. (4.11) now becomes

$$C_i(T) = C_{i,E}(K_i - K_{i,E})(T - T_E)/100 + C_i(1 + K_{i,E}(T - T_E)/100), \tag{4.17}$$

where $K_{i,E}$ represents the expected value of K_i, and (4.12) becomes

$$R_p(T) = [1 + C_E(T - T_E)/100]R_p + R_c(C - C_E)(T - T_E)/100, \tag{4.18}$$

where R_c is the center of the bin considered. Note that the effect of (4.17) and (4.18) is to replace product convolutions by convolutions of sums of random variables when it comes to computing densities. Also, supposing that $t_1 = t_N(1 - \delta)$ and $t_2 = t_N(1 + \delta)$, the limits on the innermost integral of (4.14) become $t_1/F = (1 - \delta)t_N/F$ and $t_2/F = (1 + \delta)t_N/F$, respectively.

The functional form t_N/F is to be replaced by the linearized version (4.16) with $C_1(T)$, $C_2(T)$, and $V_T^{(2)}$ substituted for C_1, C_2, and V_T, respectively. We have, therefore, after a small computation,

$$t_1/F = (1 - \delta)[R_N + A_1\Delta C_1(T) + A_2\Delta C_2(T) + A_4(\Delta V_E^{(2)} - \Delta V_E^{(1)})] \tag{4.19}$$

and, likewise,

$$t_2/F = (1 + \delta)[R_N + A_1\Delta C_1(T) + A_2\Delta C_2(T) + A_4(\Delta V_E^{(2)} - \Delta V_E^{(1)})], \tag{4.20}$$

where $\Delta C_i(T) \equiv C_i(T) - C_i$. When C_1, C_2, V, and V_T are independent, normally distributed random variables, the analysis is a bit simpler, since it is easily seen that, in this case, the pair $(R_N, \Delta V_E^{(1)})$ is bivariate normal [3, pg. 162]. In addition, one notes that (4.19) and (4.20) do not depend on C_1, C_2, and V, in the linear analysis. In Section 6, we present a numerical example following this procedure. It may be noted, by analogy with the development in paragraph 3, that the condition $t_1/F \leqslant R(T) \leqslant t_2/F$ is equivalent to requiring that $R(T)$ lie between two hyperplanes in the six-dimensional $(R_N, \Delta C_1, \Delta C_2, \Delta V_E^{(1)}, \Delta V_E^{(2)}, R(T))$ space.

5. PROBABILITY INTERVALS AT AMBIENT TEMPERATURE AFTER POTTING

When the timer is actually packaged, or potted, this procedure will produce statistical changes in the component values. These changes are known in the trade as potting shifts. Such shifts can be taken into account by convolutions of the densities previously determined with those densities evolving from the operation of potting. This has an effect on such items as the picked resistor, the capacitors, and the voltage regulator. Generally, potting shifts are represented as percentage changes from previous values, and, therefore, strictly speaking, we have another product convolution to consider. For example, we represent the value of resistance due to temperature and potting by

$$R_{\text{pot}}(T) = R_p(T)(1 + \text{CHG}_1/100), \tag{5.1}$$

where the subscript pot denotes potting and CHG_1 represents a random per cent change from the value of picked resistance at temperature. If we linearize $R_{\text{pot}}(T)$ about nominal values, we find that

$$R_{\text{pot}}(T) = (1 + \text{CHG}_{1,E}/100)R(T) + R_E(T)(\text{CHG}_1 - \text{CHG}_{1,E})/100, \tag{5.2}$$

where $R_E(T)$ is the expected value of picked resistance at temperature for the given bin and $\text{CHG}_{1,E}$ is the expected value of CHG_1. From (4.18), this is given, to a first approximation, by

$$R_E(T) = [1 + C_N(T - T_N)/100]R_c, \tag{5.3}$$

where, as before, R_c is the center of the bin interval. As for the capacitances, we assume a form

$$C_{i,\text{pot}}(T) = C_i(T)(1 + \text{CHG}_2/100), \tag{5.4}$$

so that we would linearize $C_{i,\text{pot}}(T)$ about nominal values in a manner analogous to that for $R_{\text{pot}}(T)$. Lastly, the voltage regulator value after potting is representable by

$$V_{\text{pot}} = V + \text{CHG}_3. \tag{5.5}$$

Hence, we need only go back through our analysis with $R_p(T)$ replaced by $R_{\text{pot}}(T)$, $C_i(T)$ replaced by $C_{i,\text{pot}}(T)$, and V replaced by V_{pot}. It is assumed that V_T, the cold cathode diode tube firing voltage, is unaffected by potting. One more integration, corresponding to CHG_3, is introduced in order to take account of the change in regulator voltage due to potting.

6. NUMERICAL RESULTS

Using a CDC 6600 computer, we were able to develop a computer code which can be used to predict efficiently the performance of the timer under the linearity assumptions outlined in the two previous paragraphs. The integration scheme developed will, in this paragraph, be discussed in some detail. A listing of the computer code used can be provided on request.

First of all, in (4.18), we assume that R_p has a uniform distribution across the bin which is being considered and that C is normally distributed. Let us suppose, as an example, that $C_E = -0.0235$, $T_E = 75°\text{F}$, and $\sigma(C) = 0.0078$. Then, of course, from (4.18),

$$R_p(T) = [1 - 0.0235(T - 75)/100]R_p + R_c(C + 0.0235)(T - 75)/100. \tag{6.1}$$

Therefore, $R_p(T)$ is a sum of two independent random variables, one of which is uniform and the other of which is normal and of mean 0. It follows that

$$p(R_p(T)) = \frac{1}{\sqrt{2\pi}\,(0.000078)|T - 75|R_c(r_{i+1} - r_i)[1 - 0.000235(T - 75)]} \cdot \int_{(1-0.000235(T-75))r_i}^{(1-0.000235(T-75))r_{i+1}} e^{-\frac{1}{2}\left[\frac{R_p(T)-u}{R_c(0.000078)(T-75)}\right]^2} du. \tag{6.2}$$

Letting

$$v = (u - R_p(T))/R_c(0.000078)|T - 75|,$$

(6.2) is converted into

$$p(R_p(T)) = \frac{1}{\sqrt{2\pi}(r_{i+1} - r_1)[1 - 0.000235(T - 75)]} \int_{v_1}^{v_2} e^{-\frac{1}{2}v^2} dv, \tag{6.3}$$

where

$$v_1 = [(1 - 0.000235(T - 75))r_i - R_p(T)]/R_c(0.000078)|T - 75| \tag{6.4}$$

and

$$v_2 = [(1 - 0.000235(T - 75))r_{i+1} - R_p(T)]/R_c(0.000078)|T - 75|. \tag{6.5}$$

Several cases now arise according to the value of $R_p(T)$ and according to whether or not $T \geqslant 75°\text{F}$. We first consider the case when $T \geqslant 75°$. Let us develop an inequality which allows us to assert that $v_1 \leqslant -3$. In fact, suppose that

$$R_p(T) - r_i \geqslant k_1 r_i(0.000235)(T - 75), \tag{6.6}$$

where k_1 is to be so determined that $v_1 \leqslant -3$ is valid. Upon substituting (6.6) into (6.4), one has

$$v_1 \leqslant -(k_1 + 1)r_i(0.000235)/R_c(0.000078). \tag{6.7}$$

Remembering that $r_i/R_c = 1 - \epsilon$, we find that $k_1 = \epsilon/(1 - \epsilon)$ will yield the requisite inequality. Next let us obtain an inequality which will permit us to say that $v_2 \geqslant 3$. Suppose that

$$r_{i+1} - R_p(T) \geqslant k_2 r_{i+1}(0.000235)(T - 75). \tag{6.8}$$

Then, from (6.5), we have

$$v_2 \geqslant 3(k_2 - 1)(1 + \epsilon). \tag{6.9}$$

The right side of (6.9) equals 3 when

$$k_2 = (2 + \epsilon)/(1 + \epsilon).$$

Thus, if, for $T \geqslant 75$,

$$r_i(\epsilon) = r_i\left(1 + \frac{\epsilon}{1-\epsilon}(0.000235)(T-75)\right) \leqslant R(T) \tag{6.10}$$

$$\leqslant r_{i+1}\left(1 - \frac{2+\epsilon}{1+\epsilon}(0.000235)(T-75)\right) = r_{i+1}(\epsilon),$$

it follows from (6.3) that

$$p(R_p(T)) = \frac{1}{(r_{i+1} - r_i)[1 - 0.000235(T-75)]}. \tag{6.11}$$

Now suppose that $T < 75$. Letting $R_p(T) - r_i \geqslant k_3 r_i(0.000235)(T-75)$, it follows that

$$v_1 \leqslant 3(k_3 + 1)(1-\epsilon). \tag{6.12}$$

The right side equals -3 when $k_3 = -(2-\epsilon)/(1-\epsilon)$. Again, assuming that $r_{i+1} - R_p(T) \geqslant k_4 r_{i+1}(0.000235)(T-75)$, we have

$$v_2 \geqslant -3(1+\epsilon)(k_4 - 1), \tag{6.13}$$

which equals 3 when $k_4 = \epsilon/(1+\epsilon)$. Therefore, when $T < 75$ and

$$r_i'(\epsilon) = r_i\left(1 - \frac{2-\epsilon}{1-\epsilon}(0.000235)(T-75)\right) \leqslant R(T) \tag{6.14}$$

$$\leqslant r_{i+1}\left(1 - \frac{\epsilon}{1+\epsilon}(0.000235)(T-75)\right) = r_{i+1}'(\epsilon),$$

(6.11) is again satisfied. Next let us go back to the case when $T \geqslant 75$, but let us now require that $v_2 \leqslant -3$. We find that such is true when

$$R_p(T) \geqslant r_{i+1} - \frac{\epsilon}{1+\epsilon} r_{i+1}(0.000235)(T-75). \tag{6.15}$$

Since $v_2 \leqslant -3$ also implies that $v_1 \leqslant -3$, we can assume that $p(R(T)) \cong 0$ in this case. Likewise, one finds that $v_1 \geqslant 3$ whenever

$$R_p(T) \leqslant r_i - \frac{2-\epsilon}{1-\epsilon} r_i(0.000235)(T-75), \tag{6.16}$$

so that, in this range, $p(R_p(T)) \cong 0$, also. When $T < 75$, $v_2 \leqslant -3$ whenever

$$R_p(T) \geqslant r_{i+1} - \frac{2+\epsilon}{1+\epsilon} r_{i+1}(0.000235)(T-75), \tag{6.17}$$

and $v_1 \geqslant 3$ when

$$R_p(T) \leqslant r_i + \frac{\epsilon}{1-\epsilon} r_i(0.000235)(T-75). \tag{6.18}$$

Again it follows that $p(R_p(T)) \cong 0$. Now there remain certain intervals in which $p(R_p(T))$ cannot be treated as constant for a given temperature. For example, it is found that, when $T \geqslant 75$ and

$$r_{i+1}(\epsilon) = r_{i+1}\left(1 - \frac{2+\epsilon}{1+\epsilon}(.000235)(T-75)\right) \tag{6.19}$$

$$\leqslant R_p(T) \leqslant r_{i+1}\left(1 - \frac{\epsilon}{1+\epsilon}(.000235)(T-75)\right) = s_{i+1}(\epsilon),$$

$-3 < v_2 < 3$ while $v_1 \leqslant -3$. Also, in the interval

$$(6.20)\qquad s_i(\epsilon) = r_i\left(1 - \frac{2-\epsilon}{1-\epsilon}\,(.000235)(T-75)\right) \leqslant R_p(T)$$
$$\leqslant r_i\left(1 + \frac{\epsilon}{1-\epsilon}\,(.000235)(T-75)\right) = r_i(\epsilon),$$

$-3 < v_1 < 3$ while $v_2 \geqslant 3$. When $T < 75$, $p(R_p(T))$ cannot be treated as constant whenever

$$(6.21)\qquad s_i'(\epsilon) = r_i\left(1 + \frac{\epsilon}{1-\epsilon}\,(.000235)(T-75)\right) \leqslant R_p(T) \leqslant r_i'(\epsilon)$$

or

$$(6.22)\qquad r_{i+1}'(\epsilon) \leqslant R_p(T) \leqslant r_{i+1}\left(1 - \frac{2+\epsilon}{1+\epsilon}\,(.000235)(T-75)\right) = s_{i+1}'(\epsilon).$$

The intervals so developed, in which the behavior of $p(R_p(T))$ is examined, are very important in the numerical study conducted on the CDC 6600. We now set up the precise procedure used in the computer study. First of all, referring to (4.19) and (4.20), we find it a little more natural to integrate with respect to $\Delta C_1(T)$ or $\Delta C_2(T)$ first instead of $R_p(T)$. We see then that our region of integration is fully specified by

$$(6.23)\qquad f_1(\Delta C_2, R_p(T), R_N, \Delta V_E^{(1)}, \Delta V_E^{(2)}) \leqslant \Delta C_1 \leqslant f_2(\Delta C_2, R_p(T), R_N, \Delta V_E^{(1)}, \Delta V_E^{(2)})$$
$$-\infty < \Delta C_2 < +\infty$$
$$-\infty < R_p(T) < +\infty$$
$$-\infty < \Delta V_E^{(1)} < \infty$$
$$-\infty < \Delta V_E^{(2)} < \infty$$
$$r_i \leqslant R_N \leqslant r_{i+1}$$
$$T_1 \leqslant T \leqslant T_2,$$

where, for $A_1 > 0$,

$$(6.24)\qquad f_1 = \frac{1}{A_1}\left[\frac{R_p(T)}{1+\delta} - A_2\Delta C_2 - R_N - A_4(\Delta V_E^{(2)} - \Delta V_E^{(1)})\right]$$
$$f_2 = \frac{1}{A_1}\left[\frac{R_p(T)}{1-\delta} - A_2\Delta C_2 - R_N - A_4(\Delta V_E^{(2)} - \Delta V_E^{(1)})\right]$$

and the inclusion of negative values for $R_p(T)$ is merely a mathematical artifice. The density function for this process then has the following form:

$$(6.25)\qquad h_1(\Delta C_1, \Delta C_2, R_N, R_p(T), \Delta V_E^{(1)}, \Delta V_E^{(2)})$$
$$= p_1(\Delta C_1)p_2(\Delta C_2)p_3(R_p(T))p_4(R_N, \Delta V_E^{(1)})p_5(\Delta V_E^{(2)})/p_i(T_2 - T_1).$$

The densities p_1, p_2, and p_5 are all normal densities. The mass function p_3 was ascertained in (6.3). p_4 is a bivariate normal density, and p_i is the probability of being in bin i. It is easy to determine the correlation coefficient ρ for p_4. Multiplying $\Delta V_E^{(1)}$ by R_N, as given by (4.16), we have

$$(6.26)\qquad E(R_N\Delta V_E^{(1)}) = A_4\,E(\Delta V_E^{(1)})^2$$
$$= A_4\,\sigma^2(\Delta V_E^{(1)}),$$

and, since the expected value of $\Delta V_E^{(1)}$ is zero, $\operatorname{cov}(R_N, \Delta V_E^{(1)}) = E(R_N\Delta V_E^{(1)})$. It follows, using (6.26), that

$$(6.27)\qquad \rho = A_4\sigma(\Delta V_E^{(1)})/\sigma(R_N).$$

The factors in (6.25), other than p_3, are given by

$$p_1(\Delta C_1) = \frac{1}{(2\pi)^{1/2}\sigma(\Delta C_1)} \exp\left[-\frac{1}{2}\left[\frac{\Delta C_1 - E(\Delta C_1)}{\sigma(\Delta C_1)}\right]^2\right]$$

$$p_2(\Delta C_2) = \frac{1}{(2\pi)^{1/2}\sigma(\Delta C_2)} \exp\left[-\frac{1}{2}\left[\frac{\Delta C_2 - E(\Delta C_2)}{\sigma(\Delta C_2)}\right]^2\right]$$

$$p_4(R_N, \Delta V_E^{(1)}) = \frac{1}{2\pi\sigma(R_N)\sigma(\Delta V_E^{(1)})\sqrt{1-\rho^2}}$$
$$\cdot \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left[\frac{R_N - E(R_N)}{\sigma(R_N)}\right]^2 + \left[\frac{\Delta V_E^{(1)} - E(\Delta V_E^{(1)})}{\sigma(\Delta V_E^{(1)})}\right]^2 - 2\rho\left[\frac{R_N - E(R_N)}{\sigma(R_N)}\right]\left[\frac{\Delta V_E^{(1)} - E(\Delta V_E^{(1)})}{\sigma(\Delta V_E^{(1)})}\right]\right]\right\}$$

$$p_5(\Delta V_E^{(2)}) = \frac{1}{(2\pi)^{1/2}\sigma(\Delta V_E^{(2)})} \exp\left[-\frac{1}{2}\left[\frac{\Delta V_E^{(2)} - E(\Delta V_E^{(2)})}{\sigma(\Delta V_E^{(2)})}\right]^2\right],$$

where ρ is given by (6.27) and p_4 is the well-known joint normal density for two variates [7, pp. 111-114].

Now let $K_{i,E} = .04$ for $i = 1, 2$ in (4.17) and $\sigma(K_i) = .013$, $i = 1, 2$. Recall from our discussion in paragraph IV that $C_{1,E} = .44\mu f$, $C_{2,E} = .15\mu f$, $E(R_N) = 40.16$ megohms, $A_1 = 8.65$, $A_2 = -293.12$, $A_4 = -0.95$, and $\sigma(R_N) = 4.04$. In addition, suppose that $E(\Delta V_E^{(1)}) = E(\Delta V_E^{(2)}) = 0$ and $\sigma(\Delta V_E^{(1)}) = \sigma(\Delta V_E^{(2)}) = 0.2357$. Then it is seen that

$$E(\Delta C_1) = 0.000176(T-75),\ \sigma(\Delta C_1) = 0.00005874|T-75|,$$

$$E(\Delta C_2) = 0.00006(T-75),\ \text{and}\ \sigma(\Delta C_2) = 0.00002034|T-75|.$$

Next we make several changes of variable. Let

(6.28)
$$u = (\Delta C_1 - E(\Delta C_1))/\sqrt{2}\sigma(\Delta C_1)$$
$$w = (\Delta C_2 - E(\Delta C_2))/\sqrt{2}\sigma(\Delta C_2)$$
$$z = v/\sqrt{2}$$
$$u_1 = (R_N - E(R_N))/\sqrt{2(1-\rho^2)}\sigma(R_N)$$
$$w_1 = \Delta V_E^{(1)}/\sqrt{2(1-\rho^2)}\,\sigma(\Delta V_E^{(1)})$$
$$w_2 = \Delta V_E^{(2)}/\sqrt{2}\sigma(\Delta V_E^{(2)}).$$

Then (6.25) becomes

(6.29)
$$h_2 = \frac{\sqrt{1-\rho^2}}{\pi^3(r_{i+1} - r_i)[1 + C_N(T-75)/100]} e^{-u^2} \cdot e^{-w^2} \int_{z_1}^{z_2} e^{-z^2}\, dz$$
$$\cdot e^{-w_2^2} \cdot e^{-(u_1^2 - 2\rho u_1 w_1 + w_1^2)}/p_i(T_2 - T_1).$$

Now one finds, by completing the square, that

(6.30)
$$e^{-(u_1^2 - 2\rho u_1 w_1 + w_1^2)} = e^{-(w_1 - \rho u_1)^2 - (1-\rho^2)u_1^2}.$$

Next we let $w_3 = w_1 - \rho u_1$ and $u_2 = \sqrt{1-\rho^2}\,u_1$. Our integrand becomes

$$(6.31)\qquad h = \frac{1}{\pi^3(r_{i+1}-r_i)[1+C_N(T-75)/100]}\, e^{-u^2}\cdot e^{-w^2}\int_{z_1}^{z_2} e^{-z^2}\,dz \cdot e^{-w_3^2}\cdot e^{-u_2^2}\cdot e^{-w_2^2}/p_i(T_2-T_1).$$

For brevity, set $Y=(w, w_2, w_3)$, and let $R^3(-\infty,\infty)$ denote the usual three-dimensional Euclidean space. Also, put $u_{2,i}=(r_i-E(R_N))/\sqrt{2}\,\sigma(R_N)$ and $u_{2,i+1}=(r_{i+1}-E(R_N))/\sqrt{2}\,\sigma(R_N)$. Then our integration scheme becomes

$$(6.32)\qquad P_i(t_N(1-\delta)\leqslant t\leqslant t_N(1+\delta)) = A_1 + A_2,$$

where

$$(6.33)\quad A_1=\int_{T_1}^{75}\int_{u_{2,i}}^{u_{2,i+1}}\int_{Y\in R^3(-\infty,\infty)} dY\,du_2\,dT\left[\int_{s_i'(\epsilon)}^{r_i'(\epsilon)}\int_{F_1(Y,u_2,R)}^{F_2(Y,u_2,R)} h(u,Y,R,u_2)\,du\,dR + \int_{r_i'(\epsilon)}^{r_{i+1}'(\epsilon)}\int_{F_1(Y,u_2,R)}^{F_2(Y,u_2,R)} h(u,Y,R,u_2)\,du\,dR + \int_{r_{i+1}'(\epsilon)}^{s_{i+1}'(\epsilon)}\int_{F_1(Y,u_2,R)}^{F_2(Y,u_2,R)} h(u,Y,R,u_2)\,du\,dR\right]$$

and A_2 is obtained by using 75 and T_2 for limits on the T integration in place of T_1 and 75, respectively, with primed quantities replaced by unprimed quantities. In addition, we have set

$$(6.34)\quad F_1(Y,u_2,R)=F_1(w,u_2,w_2,w_3,R)=[f_1(\Delta C_2,R,\Delta V_E^{(2)},\Delta V_E^{(1)},R_N)-E(\Delta C_1)]/\sqrt{2}\sigma(\Delta C_1)$$
$$F_2(Y,u_2,R)=F_2(w,u_2,w_2,w_3,R)=[f_2(\Delta C_2,R,\Delta V_E^{(2)},\Delta V_E^{(1)},R_N)-E(\Delta C_1)]/\sqrt{2}\sigma(\Delta C_1).$$

Now f_1 and f_2 were defined in (6.24), and, from the changes of variable given by (6.28), we have

$$(6.35)\qquad \Delta C_2 = E(\Delta C_2)+\sqrt{2}\,w\sigma(\Delta C_2)$$
$$\Delta V_E^{(2)} = \sqrt{2}\sigma(\Delta V_E^{(2)})\,w_2$$
$$\Delta V_E^{(1)} = \sqrt{2(1-\rho^2)}\sigma(\Delta V_E^{(1)})(w_3+\rho u_2/\sqrt{1-\rho^2})$$
$$R_N = E(R_N)+\sqrt{2}\sigma(R_N)u_2.$$

Our computer code is just the implementation of a nesting procedure, making use of Gaussian and Hermite-Gaussian quadrature routines, together with routines to evaluate the error integral [5, pp. 130-132], [6, pp. 319-330], [8], [1, pg. 924]. It turned out to be convenient and numerically accurate and timewise efficient to employ three Gauss points per integration step.

The effect of cold cathode diode firing voltage variations in this problem is more significant than that of ambient temperature departures from nominal. In our case study, for example, when $\epsilon=.01$ and $\delta=.02$, P_i was essentially 91%. With $\delta=.03$, this figure was increased to almost 100%. Results for six bins with $\epsilon=.01$ and $\delta=.03$ are given in Table 1.

TABLE 1 — *Performance of Fuze Timer for Representative Bins*

P_i	r_i	r_{i+1}	R_c
.994848	37.4435	38.2000	37.8218
.995103	38.2000	38.9717	38.5858
.995221	38.9717	39.7590	39.3653
.995414	39.7590	40.5622	40.1606
.995452	40.5622	41.3817	40.9719
.995490	41.3817	42.2176	41.7997

It is seen that the probability is essentially the same independent of the bin. Running time for this problem was approximately four seconds per bin. Indeed one would reason, as in paragraph 3, that, at least approximately, each bin should yield the same probability for firing time, given a $\delta - \epsilon$ combination. This should occur if the nonlinearities are not too severe and the distributions due to change in temperature and cold cathode diode firing voltage variations are fairly compact. This would then mean that we need only examine one bin to determine the performance of the timer, and our integration procedure could then represent a substantial time saving over a Monte Carlo simulation.

Going back to (6.32), we can also give an error bound for the part neglected in the computation of P_i. Let us illustrate in one case what is happening. For instance, we have neglected

$$\int_{75}^{T_2} \int_{u_{2,i}}^{u_{2,i+1}} \int_{s_{i+1}(\epsilon)}^{\infty} \int_{Y \in R^3(-\infty,\infty)} \int_{F_1(Y,u_2,R)}^{F_2(Y,u_2,R)} h(u,Y,R,u_2)\, du dY dR du_2 dT \tag{6.36}$$

Clearly, (6.36) is bounded above by

$$\int_{75}^{T_2} \int_{u_{2,i}}^{u_{2,i+1}} \int_{Z \in R^4(-\infty,\infty)} \int_{s_{i+1}(\epsilon)}^{\infty} h(R,Z,u_2)\, dR dZ du_2 dT, \tag{6.37}$$

where $Z = (u, Y)$. Noting that $h(u,w,R,w_3,u_2,w_2) = g(u,w,w_3,u_2,w_2)p(R)$ and that

$$\int_{T_1}^{T_2} \int_{u_{2,i}}^{u_{2,i+1}} \int_{Z \in R^4(-\infty,\infty)} g(Z,u_2)\, dZ du_2 dT = 1,$$

We need only study the behavior of the integration with respect to R. Going back to (6.37), when $s_{i+1} \leqslant R < \infty$, we know that $v_1 \leqslant v_2 \leqslant -3$. Therefore, it is easy to show that

$$p(R_p(T)) \leqslant \frac{1}{\sqrt{2\pi}}\, e^{-v_2^2/2}/R_c\, \sigma(C)\,|T - 75|/100. \tag{6.38}$$

It follows [2, pg. 149] that

$$\int_{s_{i+1}(\epsilon)}^{\infty} p(R(T))\, dR(T) \leqslant \frac{1}{\sqrt{2\pi}} \int_3^{\infty} e^{-x^2/2}\, dx \cong .00135. \tag{6.39}$$

A similar result is obtained when R is restricted to the interval $(-\infty, s_i(\epsilon))$ and $T \geqslant 75°$ or when R lies in either $(s'_{i+1}(\epsilon),\infty)$ or $(-\infty, s'_i(\epsilon))$ and $T < 75°$. The result is finally that the portion neglected is bounded above by .0027, so that we are at most off in the third decimal place.

7. THE CASE OF TWO OR MORE TIMERS

An interesting case study arises when there are two or more timers which are statistically dependent. This occurs, for example, when, after the first timer is operated, a switch closes and a second timer is started, the second one being fed by the same capacitor which fed the first timer. Let us suppose, for instance, that capacitor $C1$ in Figure 1 feeds the second fuze timer indicated in Figure 6.

At the end of operation of the first timer, switch S in Figure 6 is thrown into the position indicated, thus allowing $C1$ to begin charging up $C4$. $C5$ serves as the reference capacitor. The second timer is also governed by a simple first order differential equation, and one can show that the time is given by

$$t = \frac{R^{(1)} C_1 C_4}{C_1 + C_4}\, \ell n \left[\frac{C_1 V - C_2(V_T - V)}{C_1 V - C_2(V_T - V) - (V_{T,1} - V)(C_1 + C_4)} \right]. \tag{7.1}$$

Letting $t_N^{(1)}$ be the nominal time for the second timer, we find the nominal resistance for this timer to be

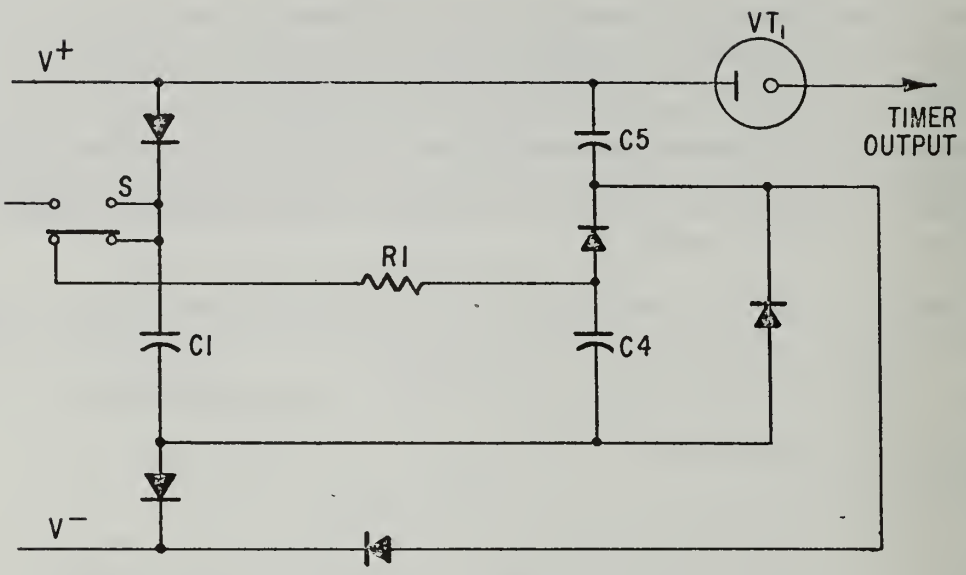


FIGURE 6. Second fuze timer configuration

$$(7.2) \qquad R_N^{(1)} = \frac{(C_1 + C_4)t_N^{(1)}}{C_1 C_4}\left\{\ell n\left[\frac{VC_1 - (V_T^{(1)} - V)C_2}{VC_1 - (V_T^{(1)} - V)C_2 - (V_{T,1}^{(1)} - V)(C_1 + C_4)}\right]\right\}^{-1},$$

where $V_T^{(1)} = V_T^E + \Delta V_E^{(1)}$ and $V_{T,1}^{(1)} = V_{T,1}^E + \Delta V_{E,1}^{(1)}$. Then, substituting (4.8) into (7.2), we derive the functional relationship

$$(7.3) \qquad R_N^{(1)} = R_N^{(1)}\,(C_1, C_2, C_4, V, R_N, \Delta V_E^{(1)}, V_{T,1}^{(1)}).$$

Solving (7.3) for $V_{T,1}^E$, we have

$$(7.4) \qquad V_{T,1}^E = h(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}).$$

To determine the joint density for the process, we must, by analogy with the method in paragraph 4, introduce a pair of diode firing variations $\Delta V_E^{(2)}$ and $\Delta V_{E,1}^{(2)}$. We then consider the following transformation of variables:

$$(7.5) \qquad \begin{aligned} V_{T,1}^E &= h(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}) \\ C_1 &= C_1 \\ C_2 &= C_2 \\ C_4 &= C_4 \\ V &= V \\ R_N &= R_N \\ \Delta V_E^{(1)} &= \Delta V_E^{(1)} \\ \Delta V_{E,1}^{(1)} &= \Delta V_{E,1}^{(1)} \\ \Delta V_E^{(2)} &= \Delta V_E^{(2)} \\ \Delta V_{E,1}^{(2)} &= \Delta V_{E,1}^{(2)}. \end{aligned}$$

To compute the density, we employ (4.10) and the Jacobian of the transformation (7.5) to obtain

$$(7.6)\quad d_5(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)})$$
$$= f(C_1, C_2, V, \Delta V_E^{(1)}, \Delta V_E^{(2)}, R_N) \cdot d_1(C_4) \cdot d_2(\Delta V_{E,1}^{(1)}) \cdot d_3(\Delta V_{E,1}^{(2)})$$
$$\cdot d_4(V_{T,1}^E) \cdot \left| \frac{\partial V_{T,1}^E}{\partial R_N^{(1)}} \right|.$$

Also, if both $R_N^{(1)}$ and R_N are linearized about nominal values of capacitance, tube firing voltages, and regulator voltage, then the map

$$(7.7)\quad R_N^{(1)} = L_1(C_1, C_2, C_4, V_T^E, \Delta V_E^{(1)}, V, V_{T,1}^E, \Delta V_{E,1}^{(1)})$$
$$R_N = L_2(C_1, C_2, V, V_T^E, \Delta V_E^{(1)})$$
$$\Delta V_{E,1}^{(1)} = \Delta V_{E,1}^{(1)}$$
$$\Delta V_E^{(1)} = \Delta V_E^{(1)}$$

shows that $(R_N^{(1)}, R_N, \Delta V_{E,1}^{(1)}, \Delta V_E^{(1)})$ is a quadrivariate normal random vector [2, pg. 162]. The reason is that all random variables on the right side of (7.7) are independent and normally distributed. At the nominal temperature, the density function is therefore generally representable by

$$(7.8)\quad d_6(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, R_p, R_p^{(1)}, \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)})$$
$$= d_5(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)})$$
$$\cdot p(R_p) \cdot p^{(1)}(R_p^{(1)}),$$

where, for example,

$$p(R_p) = 1/(r_{i+1} - r_i)$$

and

$$p^{(1)}(R_p^{(1)}) = 1/(r_{j+1}^{(1)} - r_j^{(1)})$$

if picked resistance is equally likely across the bins. (7.8), also, obviously indicates that picked resistances are statistically independent of the other component values. It will be possible to reduce (7.8) to the simpler form

$$(7.9)\quad d_6(R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, R_p, R_p^{(1)}, \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)})$$
$$= p(R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}) \cdot p(R_p) \cdot p^{(1)}(R_P^{(1)}) \cdot p(\Delta V_E^{(2)}) \cdot p(\Delta V_{E,1}^{(2)})$$

when (7.7) is valid, $p(R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)})$ being the density for the quadrivariate normal distribution [9, pg. 88]. From (7.7) the elements of the covariance matrix [9, pg. 88] can be easily obtained.

Next account must be taken of changes in component values due to temperature changes from the nominal value. We use the same ideas presented in paragraph 4, together with the same notation. The density becomes

$$(7.10)\quad d(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, R_p(T), R_p^{(1)}(T), \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)},$$
$$\Delta CP_1(T), \Delta CP_2(T), \Delta CP_4(T))$$
$$= d_5(C_1, C_2, C_4, V, R_N, R_N^{(1)}, \Delta V_E^{(1)}, \Delta V_{E,1}^{(1)}, \Delta V_E^{(2)}, \Delta V_{E,1}^{(2)})$$
$$\cdot p(\Delta CP_1(T)) \cdot p(\Delta CP_2(T)) \cdot d(\Delta CP_4(T)) \cdot p(R_p(T)) \cdot p^{(1)}(R_p^{(1)}(T)),$$

where $p(R_p(T))$ and $p^{(1)}(R_p^{(1)}(T))$ are again convolution densities. We must now determine limits of integration. One requires that the first timer fire in time t, where $t_1 = t_N(1-\delta) \leqslant t \leqslant t_N(1+\delta) = t_2$ and that the second timer fire in time $t^{(1)}$, where $t_1^{(1)} = t_N^{(1)}(1-\delta^{(1)}) \leqslant t^{(1)} \leqslant t_N^{(1)}(1+\delta^{(1)}) = t_2^{(1)}$. Therefore, we have

(7.11)
$$
\begin{aligned}
& t_1/F \leqslant R_p(T) \leqslant t_2/F \\
& t_1^{(1)}/F^{(1)} \leqslant R_p^{(1)}(T) \leqslant t_2^{(1)}/F^{(1)} \\
& -\infty < C_i < \infty, \quad i = 1, 2, 4 \\
& -\infty < \Delta CP_i(T) < \infty, \quad i = 1, 2, 4 \\
& -\infty < \Delta V_E^{(1)} < \infty \\
& -\infty < \Delta V_{E,1}^{(1)} < \infty \\
& -\infty < \Delta V_E^{(2)} < \infty \\
& -\infty < \Delta V_{E,1}^{(2)} < \infty \\
& r_i \leqslant R_N < r_{i+1} \\
& r_j^{(1)} \leqslant R_N^{(1)} < r_{j+1}^{(1)} \\
& -\infty < V < \infty,
\end{aligned}
$$

where $F = F(C_1(T), C_2(T), V, V_T^{(2)})$, $F^{(1)} = F^{(1)}(C_1(T), C_2(T), C_4(T), V, V_T^{(2)}, V_{T,1}^{(2)})$ and $V_T^{(2)} = V_T^E + \Delta V_E^{(2)}$, $V_{T,1}^{(2)} = V_{T,1}^E + \Delta V_{E,1}^{(2)}$, as before. V_T^E is given by (4.8) and $V_{T,1}^E$ by (7.5). Also, (4.13) holds for $i = 1, 2$, and 4.

An integration scheme patterned after (4.14) can then be recorded with $p_{ij} = \text{prob}\ (R_N \in \text{bin } i \text{ and } R_N^{(1)} \in \text{bin } j)$ in place of p_i. (4.15) would then be replaced by a double sum:

$$
\begin{aligned}
(7.12)\quad & P(t_1 \leqslant t \leqslant t_2,\ t_1^{(1)} \leqslant t^{(1)} \leqslant t_2^{(1)}) \\
& = \frac{1}{T_2 - T_1} \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} p_{ij} \int_{T_1}^{T_2} P_{ij}(t_1 \leqslant t \leqslant t_2,\ t_1^{(1)} \leqslant t^{(1)} \leqslant t_2^{(1)} \,|\, T)\, dT.
\end{aligned}
$$

Also, in the case where (7.7) is valid, $C_1 C_2$, C_4, and V are eliminated and $\Delta CP_i(T)$ is to be replaced by $\Delta C_i(T)$. In addition, $t^{(1)}/F^{(1)}$ and t/F become linear forms in $\Delta C_1(T)$, $\Delta C_2(T)$, $\Delta C_4(T)$, R_N, $R_N^{(1)}$, $\Delta V_E^{(1)}$, $\Delta V_E^{(2)}$, $\Delta V_{E,1}^{(1)}$, and $\Delta V_{E,1}^{(2)}$. In that case a sixteen-fold integral is reduced to a twelve-fold integral.

ACKNOWLEDGMENTS

The author would like to thank personally Messrs. Larry Burkhardt and John Abell, of the fuze group at this laboratory, for their generous support of this work and their helpful suggestions. Also, he would like to thank Mr. William McDonald and Mr. Ted Orlow, also of this laboratory, for their helpful comments and guiding ideas.

REFERENCES

[1] Abramowitz, Milton and Irene A. Stegun (Editors), Handbook of Mathematical Functions, National Bureau of Standards Applied Mathematics Series, No. 55, June (1964).

[2] Cohen, Edgar A., Jr. and Ronald Goldstein, "A Component Reliability Model for Bomb Fuze MK 344 Mod 1 and MK 376 Mod 0", NSWC/WOL/TR 75-123 (1975).

[3] Fisz, Marek, *Probability Theory and Mathematical Statistics*, John Wiley & Sons, Inc., New York (1963).
[4] Hald, A., *Statistical Theory with Engineering Applications*, John Wiley & Sons, Inc., New York, Chapman & Hall, Limited, London (1952).
[5] Hamming, R.W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York (1962).
[6] Hildebrand, F.B., *Introduction to Numerical Analysis*, McGraw-Hill, New York (1956).
[7] Hogg, Robert V. and Allen T. Craig, *Introduction to Mathematical Statistics*, 3rd Edition, MacMillan, London (1970).
[8] IBM 7094/7094 Operating System Version 13, IBJOB Processor, Appendix H: FORTRAN IV Mathematics Subroutines, International Business Machines Corporation, New York (1965).
[9] Parzen, Emanuel, *Stochastic Processes*, Holden-Day, San Francisco, London; Amsterdam (1964).

THE ASYMPTOTIC SUFFICIENCY OF SPARSE ORDER STATISTICS IN TESTS OF FIT WITH NUISANCE PARAMETERS*

Lionel Weiss

Cornell University
Ithaca, New York

ABSTRACT

In an earlier paper, it was shown that for the problem of testing that a sample comes from a completely specified distribution, a relatively small number of order statistics is asymptotically sufficient, and for all asymptotic probability calculations the joint distribution of these order statistics can be assumed to be normal. In the present paper, these results are extended to certain cases where the problem is to test the hypothesis that a sample comes from a distribution which is a member of a specified parametric family of distributions, with the parameters unspecified.

1. INTRODUCTION

For each n, the random variables $X_1(n), \ldots, X_n(n)$ are independent, identically distributed, with unknown common probability density function and cumulative distribution function $f_n(x)$, $F_n(x)$ respectively. An m-parameter family of distributions, with pdf $f_0(x;\theta_1, \ldots, \theta_m)$ and cdf $F_0(x;\theta_1, \ldots, \theta_m)$, is specified, and the problem is to test the hypothesis that $f_n(x) = f_0(x;\theta_1, \ldots, \theta_m)$ for all x, for some unspecified values of $\theta_1, \ldots, \theta_m$.

In [5] the simpler problem of testing the hypothesis that $f_n(x) = f_0(x)$, where $f_0(x)$ is completely specified, was discussed. In this simpler case, the familiar probability integral transformation can be used to reduce the problem to that of testing whether a sample comes from a uniform distribution over (0,1). This type of reduction is not always available when the hypothetical density is not completely specified. (See [1] for some cases where the reduction is available.)

Since we will be interested in large sample theory, to keep the alternatives challenging we will assume that $f_n(x) = f_0(x;\theta_1^0, \ldots, \theta_m^0)\ (1 + r_n(x))$ for some unknown values $\theta_1^0, \ldots, \theta_m^0$ and some unknown function $r_n(x)$ satisfying the conditions $\sup_x |r_n(x)| < n^{-\epsilon}$ and $\sup_x \left|\frac{d^j r_n(x)}{dx^j}\right| < n^{-\epsilon}$ for all n and for $j = 1,2,3,4$, where ϵ is a fixed value in the open interval $\left(\frac{1}{3}, \frac{1}{2}\right)$.

*Research supported by NSF Grant No. MCS76-06340.

The case where $m = 2$, and θ_1, θ_2 are location and scale parameters respectively is relatively simple to analyze, and occurs often in practice, so until Section 5 we will discuss only this case. That is, $f_0(x;\theta_1,\theta_2) = \frac{1}{\theta_2} g\left(\frac{x-\theta_1}{\theta_2}\right)$ with $\theta_2 > 0$, and the pdf $g(x)$ is completely specified. $G(x)$ denotes $\int_{-\infty}^{x} g(t)\,dt$. We assume that $\sup_x \left|\frac{d^j}{dx^j} g(x)\right| < \Delta_1 < \infty$ for $j = 1,2,3,4$, and that $\sup_x g(x) < \Delta_2 < \infty$.

For each n, we choose positive quantities p_n, q_n, and L_n satisfying the following conditions:

(1.1) $p_n < q_n < 1 - n^{-\epsilon}$.

(1.2) np_n, nq_n, L_n, and $K_n \equiv \frac{n(q_n - p_n)}{L_n}$ are all integers.

(1.3) $\lim_{n\to\infty} \frac{n^{\frac{2}{3}+\delta}}{L_n} = 1$ for some fixed δ in the open interval $\left(0, \frac{\epsilon}{2} - \frac{1}{6}\right)$.

(1.4) $\lim_{n\to\infty} p_n = 0$, $\lim_{n\to\infty} q_n = 1$, $\lim_{n\to\infty} np_n = \infty$.

(1.5) $b_n \equiv \inf_x \left\{ g(x) : G^{-1}\left(\frac{p_n}{1+n^{-\epsilon}}\right) \leqq x \leqq G^{-1}\left(\frac{q_n}{1-n^{-\epsilon}}\right) \right\} \geqq n^{-\gamma}$ for a fixed positive γ with $\frac{1}{3} - \epsilon + 2\delta + 5\gamma < 0$.

(1.6) $\lim_{n\to\infty} \frac{n^{2\epsilon}}{np_n} = \infty$, $\lim_{n\to\infty} \frac{n^{2\epsilon}}{n(1-q_n)} = \infty$.

(1.7) $\frac{g(G^{-1}(p_n))}{g(x)} > \Delta_3 > 0$ for all $x < G^{-1}(p_n)$, and
$\frac{g(G^{-1}(q_n))}{g(x)} > \Delta_4 > 0$ for all $x > G^{-1}(q_n)$.

$Y_1(n) < Y_2(n) < \cdots < Y_n(n)$ denote the ordered values of $X_1(n), \ldots, X_n(n)$. For typographical simplicity, we denote $Y_j(n)$ by Y_j. For $j = 1, \ldots, K_n$, let $\bar{Y}_j(n)$ denote $\frac{1}{2}(Y_{np_n+jL_n} + Y_{np_n+(j-1)L_n})$, and let $D_j(n)$ denote $(Y_{np_n+jL_n} - Y_{np_n+(j-1)L_n})$. For $j = 1, \ldots, K_n - 1$, let $W'(1,j,n), \ldots, W'(L_n-1,j,n)$ denote the values of the $L_n - 1$ variables among $\{X_1(n), \ldots, X_n(n)\}$ which fall in the open interval $\left\{\bar{Y}_j(n) - \frac{D_j(n)}{2}, \bar{Y}_j(n) + \frac{D_j(n)}{2}\right\}$, written in random order: that is, the same order in which the corresponding elements of $\{X_1(n), \ldots, X_n(n)\}$ are written. Define $W(i,j,n)$ as $\frac{W'(i,j,n) - \bar{Y}_j(n)}{D_j(n)}$ for $i =$

$1, \ldots, L_n - 1$ and $j = 1, \ldots, K_n$, so $-\frac{1}{2} \leqq W(i,j,n) \leqq \frac{1}{2}$. Let $\underset{\sim}{W}(j,n)$ denote the $(L_n - 1)$-dimensional vector $\{W(1,j,n), \ldots, W(L_n - 1,j,n)\}$ for $j = 1, \ldots, K_n$. Let $W(1,0,n), \ldots, W(np_n - 1,0,n)$ denote the values of the $np_n - 1$ variables among $\{X_1(n), \ldots, X_n(n)\}$ which fall in the open interval $(-\infty, Y_{np_n})$ written in random order. Let $\underset{\sim}{W}(0,n)$ denote the vector $\{W(1,0,n), \ldots, W(np_n - 1,0,n)\}$. Let $W(1,K_n+1,n), \ldots, W(n - nq_n, K_n+1,n)$ denote the values of the $n - nq_n$ variables among $\{X_1(n), \ldots, X_n(n)\}$ which fall in the open interval (Y_{nq_n}, ∞), written in random order. Let $\underset{\sim}{W}(K_n+1,n)$ denote the vector $\{W(1,K_n+1,n), \ldots, W(n - nq_n, K_n+1,n)\}$. Let $\underset{\sim}{T}(n)$ denote the (K_n+1)-dimensional vector $\{Y_{np_n+jL_n}; j = 0, 1, \ldots, K_n\}$. Note that if we are given the $K_n + 3$ vectors defined, we can compute the n order statistics $Y_1, \ldots, Y_n$, so that any test procedure based on the order statistics can also be based on the $K_n + 3$ vectors.

Let $h_n(\underset{\sim}{t}(n))$ denote the joint pdf for the elements of the vector $\underset{\sim}{T}(n)$, and let $h^*_{i,n}(\underset{\sim}{w}(i,n) \mid \underset{\sim}{t}(n))$ denote the joint conditional pdf for the elements of the vector $\underset{\sim}{W}(i,n)$, given that $\underset{\sim}{T}(n) = \underset{\sim}{t}(n)$. Then the joint pdf for all n elements of all the vectors is $h_n(\underset{\sim}{t}(n)) \prod_{i=0}^{K_n+2} h^*_{i,n}(\underset{\sim}{w}(i,n) \mid \underset{\sim}{t}(n))$, which we denote by $h_n^{(1)}$.

Next we construct two different "artificial" joint pdfs for the n elements of the vectors.

In the first artificial joint pdf, the marginal pdf for $\underset{\sim}{T}(n)$ and the conditional pdfs for $\underset{\sim}{W}(0,n)$ and $\underset{\sim}{W}(K_n+1,n)$ are the same as above. The pdfs for the elements of the other vectors are constructed as follows.

Let $\alpha_j(n)$ denote $G^{-1}\left(\frac{np_n + \left(j - \frac{1}{2}\right)L_n}{n}\right)$, and $\gamma_j(n)$ denote $\frac{L_n}{2n} \frac{g'(\alpha_j(n))}{g^2(\alpha_j(n))}$, for $j = 1, \ldots, K_n$. Let $U(i,j)$ $(i = 1, \ldots, L_n - 1; j = 1, \ldots, K_n)$ be IID random variables, independent of $\underset{\sim}{T}(n)$, $\underset{\sim}{W}(0,n)$, $\underset{\sim}{W}(K_n+1,n)$, and each with a uniform distribution over (0,1). Then the distribution of $W(i,j,n)$ is to be the distribution of $-\frac{1}{2} + (1+\gamma_j(n))U(i,j) - \gamma_j(n)U^2(i,j)$, for $i = 1, \ldots, L_n - 1$ and $j = 1, \ldots, K_n$. Denote the resulting joint pdf for all n elements by $h_n^{(2)}$.

In the second artificial joint distribution, the marginal pdf for $\underset{\sim}{T}(n)$ and the conditional pdfs for $\underset{\sim}{W}(1,n), \ldots, \underset{\sim}{W}(K_n,n)$ given $\underset{\sim}{T}(n)$ are the same as in $h_n^{(1)}$. Given $\underset{\sim}{T}(n)$, the $np_n - 1$ elements of $\underset{\sim}{W}(0,n)$ are distributed as IID random variables, each with pdf $g((x-\theta_1^0)/\theta_2^0)/\theta_2^0\, G((Y_{np_n} - \theta_1^0)/\theta_2^0)$ for $x < Y_{np_n}$, zero if $x > Y_{np_n}$. Given $\underset{\sim}{T}(n)$, the $n - nq_n$ elements of $\underset{\sim}{W}(K_n+1,n)$ are distributed as IID random variables, each with pdf $((1/\theta_2^0)\, g((x - \theta_1^0)/\theta_2^0)/(1 - G((Y_{nq_n} - \theta_1^0)/\theta_2^0))$ for $x > Y_{nq_n}$, zero if $x < Y_{nq_n}$. Denote the resulting joint pdf for all n elements by $h_n^{(3)}$.

If S_n is any measurable region in n-dimensional space, let $P_{h_n^{(i)}}(S_n)$ denote the probability assigned to S_n by the pdf $h_n^{(i)}$. The next two sections are devoted to proving the following:

THEOREM 1: $\lim_{n\to\infty} \sup_{S_n} |P_{h_n^{(2)}}(S_n) - P_{h_n^{(1)}}(S_n)| = 0.$

THEOREM 2: $\lim_{n\to\infty} \sup_{S_n} |P_{h_n^{(3)}}(S_n) - P_{h_n^{(1)}}(S_n)| = 0.$

2. PROOF OF THEOREM 1

Let $h_n^{(4)}$ denote the joint pdf which differs from $h_n^{(2)}$ only in that $\gamma_j(n)$ is replaced by $\bar{\gamma}_j(n)$, defined as $\frac{L_n}{2n}\frac{f_n'(\bar{\alpha}_j(n))}{f_n^2(\bar{\alpha}_j(n))}$, where $\bar{\alpha}_j(n) = F_n^{-1}\left(\frac{np_n + \left[j-\frac{1}{2}\right]L_n}{n}\right)$. It was shown in [8] that $\lim_{n\to\infty}\sup_{S_n}|P_{h_n^{(4)}}(S_n) - P_{h_n^{(1)}}(S_n)| = 0$, and thus Theorem 1 will be proved if we can show that $\lim_{n\to\infty}\sup_{S_n}|P_{h_n^{(2)}}(S_n) - P_{h_n^{(4)}}(S_n)| = 0$. By the reasoning used in [8], this last equality will be demonstrated if we can show that

$$\log\frac{h_n^{(2)}(\underset{\sim}{T}(n),\ \underset{\sim}{W}(0,n),\ \ldots,\ \underset{\sim}{W}(K_n+1,n))}{h_n^{(4)}(\underset{\sim}{T}(n),\ \underset{\sim}{W}(0,n),\ \ldots,\ \underset{\sim}{W}(K_n+1,n))} \equiv R_n,$$

say, converges stochastically to zero as n increases, when the joint pdf is actually $h_n^{(2)}$. From the definitions above, and the formula in [8], for all sufficiently large n we can write R_n as

$$\frac{1}{2}\sum_{j=1}^{K_n}\sum_{i=1}^{L_n-1}\left\{\begin{matrix}\log[1+\bar{\gamma}_j^2(n)-4\bar{\gamma}_j(n)W(i,j,n)] \\ -\log[1+\gamma_j^2(n)-4\gamma_j(n)W(i,j,n)]\end{matrix}\right\} \tag{2.1}$$

where $W(i,j,n)$ have the same distribution as $-\frac{1}{2} + (1+\gamma_j(n))U(i,j) - \gamma_j(n)U^2(i,j)$. We show that the expression (2.1) converges stochastically to zero as n increases by means of three lemmas. (The order symbol $0(\ )$ used below has the usual interpretation.)

LEMMA 2.1: $\max_{1\le j\le K_n}|\gamma_j(n)| = 0\left(n^{-\frac{1}{3}+\delta+2\gamma}\right)$.

PROOF: Directly from the assumptions and the definition of $\gamma_j(n)$.

LEMMA 2.2: $\sup_{p_n\le t\le q_n}|F_n^{-1}(t) - \{\theta_1^0+\theta_2^0 G^{-1}(t)\}| = 0(n^{-\epsilon+\gamma})$.

PROOF: Since $f_n(x) = \frac{1}{\theta_2^0}g\left(\frac{x-\theta_1^0}{\theta_2^0}\right)(1+r_n(x))$, with $|r_n(x)| < n^{-\epsilon}$, we have $F_n(x) = G\left(\frac{x-\theta_1^0}{\theta_2^0}\right) + \bar{R}_n(x)$, where $\bar{R}_n(x) = \int_{-\infty}^{x}\frac{1}{\theta_2^0}g\left(\frac{t-\theta_1^0}{\theta_2^0}\right)r_n(t)\,dt$, and thus $|\bar{R}_n(x)| \le n^{-\epsilon}G\left(\frac{x-\theta_1^0}{\theta_2^0}\right)$. Then we can write $F_n(x) = G\left(\frac{x-\theta_1^0}{\theta_2^0}\right)(1+R_n(x))$, where $|R_n(x)| \le n^{-\epsilon}$ for all x. Fix any value t in the closed interval $[p_n,q_n]$. Writing $F_n(x) = t = G\left(\frac{x-\theta_1^0}{\theta_2^0}\right)(1+R_n(x))$, we have $x = F_n^{-1}(t)$ and $G\left(\frac{F_n^{-1}(t)-\theta_1^0}{\theta_2^0}\right) = \frac{t}{1+R_n(F_n^{-1}(t))}$, so

$$G^{-1}\left(\frac{t}{1+n^{-\epsilon}}\right) \le \frac{F_n^{-1}(t)-\theta_1^0}{\theta_2^0} \le G^{-1}\left(\frac{t}{1-n^{-\epsilon}}\right). \tag{2.2}$$

We can write $G^{-1}\left(\frac{t}{1+n^{-\epsilon}}\right) = G^{-1}(t) - \frac{tn^{-\epsilon}}{1+n^{-\epsilon}}\left(\frac{1}{g(G^{-1}(t^*))}\right)$ where t^* is in the open

interval $\left(\frac{t}{1+n^{-\epsilon}}, t\right]$, and thus $\frac{1}{g(G^{-1}(t^*))} \leqq n^{\gamma}$, by assumption (1.5). Then $\sup_{p_n \leqq t \leqq q_n} \left|G^{-1}\left(\frac{t}{1+n^{-\epsilon}}\right) - G^{-1}(t)\right| = 0(n^{-\epsilon+\gamma})$. By a completely analogous argument, it can be shown that $\sup_{p_n \leqq t \leqq q_n} \left|G^{-1}\left(\frac{t}{1-n^{-\epsilon}}\right) - G^{-1}(t)\right| = 0(n^{-\epsilon+\gamma})$. Then the lemma follows immediately, using the inequalities (2.2).

LEMMA 2.3: $\bar{\gamma}_j(n) = \gamma_j(n) + \delta_j(n)$, where $\max_{1 \leqq j \leqq K_n} |\delta_j(n)| = 0\left(n^{-\frac{1}{3}+\delta-\epsilon+3\gamma}\right)$.

PROOF: By lemma 2.2, we can write $\bar{\gamma}_j(n)$ as

$$\frac{L_n}{2n} \frac{f_n'(\theta_1^0+\theta_2^0\alpha_j(n)+\bar{\delta}_j(n))}{f_n^2(\theta_1^0+\theta_2^0\alpha_j(n)+\bar{\delta}_j(n))},$$

where $\max_{1 \leqq j \leqq K_n} |\bar{\delta}_j(n)| = 0(n^{-\epsilon+\gamma})$, $f_n'(x) = \frac{1}{\theta_2^0} g\left(\frac{x-\theta_1^0}{\theta_2^0}\right) r_n'(x) + (1+r_n(x)) \frac{1}{(\theta_2^0)^2} g'\left(\frac{x-\theta_1^0}{\theta_2^0}\right)$, so we can write $f_n'(\theta_1^0 + \theta_2^0 \alpha_j(n) + \bar{\delta}_j(n))$ as $\frac{1}{(\theta_2^0)^2} g'(\alpha_j(n)) + \delta_j^*(n)$, where $\max_{1 \leqq j \leqq K_n} |\delta_j^*(n)| = 0(n^{-\epsilon+\gamma})$. We can also write $f_n(\theta_1^0 + \theta_2^0\alpha_j(n) + \bar{\delta}_j(n))$ as $\frac{1}{\theta_2^0} g(\alpha_j(n)) + \hat{\delta}_j(n)$, and thus $f_n^2(\theta_1^0 + \theta_2^0\alpha_j(n) + \bar{\delta}_j(n))$ as $\frac{1}{(\theta_2^0)^2} g^2(\alpha_j(n)) + \hat{\delta}_j^*(n)$, where $\max_{1 \leqq j \leqq K_n} |\hat{\delta}_j(n)| = 0(n^{-\epsilon+\gamma})$ and $\max_{1 \leqq j \leqq K_n} |\hat{\delta}_j^*(n)| = 0(n^{-\epsilon+\gamma})$. Thus we can write $\bar{\gamma}_j(n)$ as $\frac{L_n}{2n}$ $\{((1/(\theta_2^0)^2) g'(\alpha_j(n)) + \delta_j^*(n))/((1/(\theta_2^0)^2) g^2(\alpha_j(n)) + \hat{\delta}_j^*(n))\}$, and the proof of the lemma follows directly from assumptions (1.3) and (1.5).

Now we complete the proof of Theorem 1 by applying the expansion $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4(1+\omega x)^4}$ for $|x| < 1$, where $|\omega| < 1$, to each of the logarithms in the expression (2.1). This enables us to write the expression (2.1) as the sum of a finite number of expressions, each of which can easily be shown to converge stochastically to zero as n increases, using the lemmas. For example, two of these expressions are:

$$\frac{1}{2} \sum_{j=1}^{K_n} \sum_{i=1}^{L_n-1} (\bar{\gamma}_j^2(n) - \gamma_j^2(n)), \text{ and} \tag{2.3}$$

$$2 \sum_{j=1}^{K_n} \sum_{i=1}^{L_n-1} (\gamma_j(n) - \bar{\gamma}_j(n)) W(i,j,n). \tag{2.4}$$

The expression (2.3) is the sum of $K_n(L_n-1)$ terms, where $K_n(L_n-1) < n$. A typical term can be written as $(\gamma_j(n)-\bar{\gamma}_j(n))$ $(\gamma_j(n)+\bar{\gamma}_j(n))$, which by Lemmas 2.1 and 2.3 is $0(n^{-\frac{2}{3}-\epsilon+2\delta+5\gamma})$. So the whole expression (2.3) is $0(n^{\frac{1}{3}-\epsilon+2\delta+5\gamma})$ and converges to zero as n increases, by assumption (1.5). The expected value of the expression (2.4) is $2 \sum_{j=1}^{K_n} \sum_{i=1}^{L_n-1} (\gamma_j(n)-\bar{\gamma}_j(n)) \left(\frac{1}{6}\right) \gamma_j(n)$, and the variance of the expression (2.4) is $4 \sum_{j-1}^{K_n} \sum_{i=1}^{L_n-1}$

$(\gamma_j(n)-\bar{\gamma}_j(n))^2\left[\frac{1}{12}+\frac{\gamma_j^2(n)}{180}\right]$. This mean and variance can both be seen to converge to zero as n increases by the same reasoning as in the analysis of the expression (2.3), and thus the expression (2.4) converges stochastically to zero as n increases. The other expressions in the sum comprising the expression (2.1) can be handled similarly, completing the proof of Theorem 1.

3. PROOF OF THEOREM 2

In Section 2 we showed that we can write $F_n(x) = G\left[\frac{x-\theta_1^0}{\theta_2^0}\right](1+R_n(x))$ where $|R_n(x)| \leqq n^{-\epsilon}$ for all x. We now develop an analogous expression for $1-F_n(x)$. $1-F_n(x) = \int_x^\infty f_n(t)\,dt = 1 - G\left[\frac{x-\theta_1^0}{\theta_2^0}\right] + \int_x^\infty r_n(t)\,\frac{1}{\theta_2^0}\,g\left[\frac{t-\theta_1^0}{\theta_2^0}\right]dt$, and since $\left|\int_x^\infty r_n(t)\,\frac{1}{\theta_2^0}\,g\left[\frac{t-\theta_1^0}{\theta_2^0}\right]dt\right| \leqq n^{-\epsilon}\left[1-G\left[\frac{x-\theta_1^0}{\theta_2^0}\right]\right]$, we can write $1-F_n(x) = \left[1-G\left[\frac{x-\theta_1^0}{\theta_2^0}\right]\right](1+S_n(x))$ where $|S_n(x)| \leqq n^{-\epsilon}$ for all x.

Theorem 2 will be proved if we can show that

$$\log \frac{h_n^{(1)}(\underset{\sim}{T}(n),\ \underset{\sim}{W}(0,n),\ \ldots,\ \underset{\sim}{W}(K_n+1,n))}{h_n^{(3)}(\underset{\sim}{T}(n),\ \underset{\sim}{W}(0,n),\ \ldots,\ \underset{\sim}{W}(K_n+1,n))} \equiv R_n^*,$$

say, converges stochastically to zero as n increases, when the joint pdf is actually $h_n^{(1)}$. Assuming $h_n^{(1)}$ is the joint pdf, the conditional (given $\underset{\sim}{T}(n)$) distribution of R_n^* is the same as the distribution of $Q_n(1) + Q_n(2)$, where $Q_n(1) = \sum_{i=1}^{np_n-1} \log(1 + r_n(\bar{V}_i)) - (np_n - 1)\log(1 + R_n(Y_{np_n}))$, and $Q_n(2) = \sum_{i=1}^{n-nq_n} \log(1 + r_n(\bar{Z}_i)) - (n - nq_n)\log(1 + S_n(Y_{nq_n}))$, and $\bar{V}_1, \ldots, \bar{V}_{np_n-1}, \bar{Z}_1, \ldots, \bar{Z}_{n-nq_n}$ are mutually independent, each $\bar{V}_i$ with pdf $\frac{f_n(v)}{F_n(Y_{np_n})}$ for $v < Y_{np_n}$, zero for $v > Y_{np_n}$, each $\bar{Z}_i$ with pdf $\frac{f_n(z)}{1-F_n(Y_{nq_n})}$ for $z > Y_{nq_n}$, zero for $z < Y_{nq_n}$.

LEMMA 3.1: $Q_n(1)$ converges stochastically to zero as n increases.

PROOF: Define $\bar{Q}_n(1)$ as $\sum_{i=1}^{np_n-1} r_n(\bar{V}_i) - (np_n-1)R_n(Y_{np_n})$. By assumption 1.6 $|Q_n(1)-\bar{Q}_n(1)|$ converges stochastically to zero as n increases. Thus the lemma will be proved if we show that $\bar{Q}_n(1)$ converges stochastically to zero as n increases.

$$E\{r_n(\bar{V}_i)|\underset{\sim}{T}(n)\} = \frac{\int_{-\infty}^{Y_{np_n}} r_n(t)\,\frac{1}{\theta_2^0}\,g\left[\frac{t-\theta_1^0}{\theta_2^0}\right](1+r_n(t))\,dt}{G\left[\frac{Y_{np_n}-\theta_1^0}{\theta_2^0}\right](1+R_n(Y_{np_n}))} =$$

$$\frac{G_0\left(\frac{Y_{np_n}-\theta_1^0}{\theta_2^0}\right)R_n(Y_{np_n})+\bar{\omega}_n n^{-2\epsilon}G\left(\frac{Y_{np_n}-\theta_1^0}{\theta_2^0}\right)}{G_0\left(\frac{Y_{np_n}-\theta_1^0}{\theta_2^0}\right)(1+R_n(Y_{np_n}))}$$

where $|\bar{\omega}_n| \leqq 1$. From this, it follows that $|E\{r_n(\bar{V}_i)|\underset{\sim}{T}(n)\} - R_n(Y_{np_n})| = 0_p(n^{-2\epsilon})$. This implies that $E\{\bar{Q}_n(1)|\underset{\sim}{T}(n)\}$ converges to zero as n increases, and also that Variance $\{r_n(\bar{V}_i)|\underset{\sim}{T}(n)\} = 0_p(n^{-2\epsilon})$ which in turn implies that Variance $\{\bar{Q}_n(1)|\underset{\sim}{T}(n)\}$ converges stochastically to zero as n increases. These facts clearly imply that $\bar{Q}_n(1)$ converges stochastically to zero as n increases.

LEMMA 3.2: $Q_n(2)$ converges stochastically to zero as n increases.

PROOF: Define $\bar{Q}_n(2)$ as $\sum_{i=1}^{n-nq_n} r_n(\bar{Z}_i) - (n-nq_n)S_n(Y_{nq_n})$. Just as in Lemma 3.1, all we have to do is to prove that $\bar{Q}_n(2)$ converges stochastically to zero as n increases. $E\{r_n(\bar{Z}_i)|\underset{\sim}{T}(n)\} =$

$$\frac{\int_{Y_{nq_n}}^{\infty} r_n(t)\frac{1}{\theta_2^0}g\left(\frac{t-\theta_1^0}{\theta_2^0}\right)(1+r_n(t))\,dt}{\left[1-G\left(\frac{Y_{nq_n}-\theta_1^0}{\theta_2^0}\right)\right][1+S_n(Y_{nq_n})]} =$$

$$\frac{S_n(Y_{nq_n})\left[1-G\left(\frac{Y_{nq_n}-\theta_1^0}{\theta_2^0}\right)\right]+\hat{\omega}_n n^{-2\epsilon}\left[1-G\left(\frac{Y_{nq_n}-\theta_1^0}{\theta_2^0}\right)\right]}{\left[1-G\left(\frac{Y_{nq_n}-\theta_1^0}{\theta_2^0}\right)\right][1+S_n(Y_{nq_n})]}$$

where $|\hat{\omega}_n| \leqq 1$. From this, it follows that $|E\{r_n(\bar{Z}_i)|\underset{\sim}{T}(n)\} - S_n(Y_{nq_n})| = 0_p(n^{-2\epsilon})$. The rest of the proof is similar to the proof of Lemma 3.1.

Lemmas 3.1 and 3.2 imply that R_n^* converges stochastically to zero as n increases, and this proves Theorem 2.

4. CONSEQUENCES OF THE THEOREMS

Theorem 1 implies that a statistician who knows only the vectors $\underset{\sim}{T}(n)$, $\underset{\sim}{W}(0,n)$, $\underset{\sim}{W}(K_n+1,n)$ is asymptotically as well off as a statistician who knows all the vectors $\underset{\sim}{T}(n)$, $\underset{\sim}{W}(0,n)$, $\underset{\sim}{W}(1,n)$, ..., $\underset{\sim}{W}(K_n+1,n)$. This is so because given $\underset{\sim}{T}(n)$, using a table of random numbers it is possible to generate additional random variables so the joint distribution of the additional random variables and the elements of $\underset{\sim}{T}(n)$, $\underset{\sim}{W}(0,n)$, $\underset{\sim}{W}(K_n+1,n)$ is the joint distribution given by $h_n^{(2)}$. But Theorem 1 states that all probabilities computed using $h_n^{(2)}$ are asymptotically the same as probabilities computed under the actual pdf $h_n^{(1)}$.

Theorem 2 implies that asymptotically the order statistics $\{Y_1, \ldots, Y_{np_n-1}, Y_{nq_n+1}, \ldots, Y_n\}$ contain no information about $r_n(x)$. This is so because under $h_n^{(3)}$ the conditional distribution (given $\underset{\sim}{T}(n)$) of these order statistics does not involve $r_n(x)$.

Taken together, the two theorems imply that a knowledge of $\underset{\sim}{T}(n)$ is asymptotically as good as a knowledge of the whole sample, for the purpose of testing whether $r_n(x) = 0$. This assumes that we have to deal only with the challenging alternatives described in Section 1, but less challenging alternatives do not pose any problem asymptotically.

5. EXTENSION TO OTHER CASES

The results above were for the case where the unknown parameters are location and scale parameters. In other cases, it may not be possible to choose p_n and q_n that will guarantee that assumptions (1.5) and (1.6) hold for all $\theta_1, \ldots, \theta_m$, if we want $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} q_n = 1$. But if we fix p and q with $0 < p < q < 1$, an analogue of Theorem 1 can often be proved with p_n replaced by p, q_n replaced by q, and $\alpha_j(n)$, $\gamma_j(n)$

defined as $F_0^{-1}\left(\dfrac{np + \left(j - \frac{1}{2}\right)L_n}{n}; \hat{\theta}_1, \ldots, \hat{\theta}_m\right)$,

$$\frac{L_n}{2n} \frac{f_0'(\alpha_j(n); \hat{\theta}_1, \ldots, \hat{\theta}_m)}{f_0^2(\alpha_j(n); \hat{\theta}_1, \ldots, \hat{\theta}_m)} \text{ respectively,}$$

where $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are estimates of $\theta_1^0, \ldots, \theta_m^0$ based on $\{Y_{np}, Y_{np+L_n}, \ldots, Y_{nq}\}$. Then, if we are willing to ignore departures from the hypothesis in the tails of the distribution, we can still use only the order statistics $\{Y_{np}, Y_{np+L_n}, \ldots, Y_{nq}\}$.

6. APPLICATIONS

For the case where $m = 2$ and θ_1, θ_2 are location and scale parameters respectively, various tests based on $\underset{\sim}{T}(n)$ have been investigated in [2] and [6]. In particular, [2] contains various analogues of the familiar Wilk-Shapiro test, first proposed in [3]. The tests in [2] and [6] were based on $\underset{\sim}{T}(n)$ because it made the analysis easier. The present paper gives a theoretical justification for basing tests on these sparse order statistics alone.

For the location and scale parameter case, we can construct other tests, as follows. For $j = 0,1, \ldots, K_n$, let $V_j(n)$ denote $\sqrt{n}\, f_n\left[F_n^{-1}\left(\dfrac{np_n + jL_n}{n}\right)\right]\left[Y_{np_n+jL_n} - F_n^{-1}\left(\dfrac{np_n + jL_n}{n}\right)\right]$, and let $Z_j(n)$ denote $\sqrt{n}\, \dfrac{1}{\theta_2^0}\, g\left[G^{-1}\left(\dfrac{np_n + jL_n}{n}\right)\right]\left[Y_{np_n+jL_n} - F_n^{-1}\left(\dfrac{np_n + jL_n}{n}\right)\right]$.

It was shown in [4] that for all asymptotic probability calculations, we can assume that the joint distribution of $\{V_0(n), \ldots, V_{K_n}(n)\}$ is given by the normal pdf

$$c_n \exp\left\{-\frac{n(L_n-1)}{2L_n^2}\left[\frac{L_n v_0^2}{np_n}+\frac{L_n v_{K_n}^2}{n(1-q_n)}+\sum_{j=1}^{K_n}(v_j-v_{j-1})^2\right]\right\}.$$

Under the additional condition that $\frac{1}{2}-\frac{3\delta}{2}-\epsilon+2\gamma<0$, it can be shown that for all asymptotic probability calculations we can assume that the joint distribution of $\{Z_0(n), \ldots, Z_{K_n}(n)\}$ is given by the normal pdf just described. Then, if we define ρ_1 as $-\frac{p_n}{q_n}\sqrt{\frac{L_n}{np_n}}\left[1+\frac{1}{\sqrt{1-q_n}}\right]$, ρ_2 as $\sqrt{\frac{L_n}{np_n}}\,\rho_1$, and the observable random variables Q_0, $Q_1, \ldots, Q_{K_n}$ as

$$Q_0=\sqrt{\frac{n(L_n-1)}{L_n^2}}\left[\sqrt{\frac{L_n}{np_n}}\left(g(G^{-1}(p_n))\,Y_{np_n}+\rho_1 g(G^{-1}(q_n))\,Y_{nq_n}\right]\right.,$$

$$Q_j=\sqrt{\frac{n(L_n-1)}{L_n^2}}\left\{\begin{array}{l} g\left(G^{-1}\left(\frac{np_n+jL_n}{n}\right)\right)Y_{np_n+jL_n}+\rho_2 g(G^{-1}(q_n))\,Y_{nq_n} \\ -\,g\left(G^{-1}\left(\frac{np_n+(j-1)L_n}{n}\right)\right)Y_{np_n+(j-1)L_n}\end{array}\right\}$$

for $j=1, \ldots, K_n$, a straightforward computation shows that for all asymptotic probability calculations we can assume that $Q_0, Q_1, \ldots, Q_{K_n}$ are independent, each with a normal distribution with standard deviation θ_2^0, and with

$$E\{Q_0\}=\sqrt{\frac{n(L_n-1)}{L_n^2}}\left\{\sqrt{\frac{L_n}{np_n}}\,h_n(0)+\rho_1 h_n(K_n)\right\},$$

$$E\{Q_j\}=\sqrt{\frac{n(L_n-1)}{L_n^2}}\{h_n(j)-h_n(j-1)+\rho_2 h_n(K_n)\},\ \text{for } j=1, \ldots, K_n,$$

where $h_n(j)=g\left(G^{-1}\left(\frac{np_n+jL_n}{n}\right)\right)F_n^{-1}\left(\frac{np_n+jL_n}{n}\right)$. If the hypothesis is true, $F_n^{-1}\left(\frac{np_n+jL_n}{n}\right)=\theta_1^0+\theta_2^0 G^{-1}\left(\frac{np_n+jL_n}{n}\right)$, and in this case we can write $E\{Q_j\}=A_n(j)\theta_1^0+B_n(j)\theta_2^0$, where $A_n(j)$, $B_n(j)$ are known, for $j=0, \ldots, K_n$. So we have reduced our hypothesis testing problem to the following: we observe random variables $Q_0, Q_1, \ldots, Q_{K_n}$ which are independent and normal, each with the same standard deviation θ_2^0, which is unknown. The problem is to test the hypothesis that $E\{Q_j\}=A_n(j)\theta_1^0+B_n(j)\theta_2^0$, for some unknown θ_1^0, where $A_n(j)$ and $B_n(j)$ are known values, for $j=0,1, \ldots, K_n$, against alternatives that $E\{Q_j\}=A_n(j)\theta_1^0+B_n(j)\theta_2^0+\Delta_n(j)$, where $\Delta_n(j)$ is unknown.

The formulation of the problem just described makes it easy to construct various tests. For example, suppose for convenience that K_n+1 is a multiple of 4. Then it is possible to find $\frac{1}{4}(K_n+1)$ sets of nonrandom quantities $\left\{\lambda_n(4i), \lambda_n(4i+1), \lambda_n(4i+2), \lambda_n(4i+3); i=0, \ldots, \frac{K_n-3}{4}\right\}$ such that the $\frac{1}{4}(K_n+1)$ quantities $\bar{Q}_n(i)=\lambda_n(4i)Q_i+\lambda_n(4i+1)Q_{i+1}+$

$\lambda_n(4i+2)Q_{i+2} + \lambda_n(4i+3)Q_{i+3}$ $\left[i = 0, 1, \ldots, \frac{K_n-3}{4}\right]$ can be assumed to be independent normal random variables, each with unknown standard deviation θ_2^0, and with $E\{\bar{Q}_n(i)\} = \sum_{j=0}^{3} \lambda_n(4i+j)\Delta_n(4i+j) = \bar{\Delta}_n(i)$, say, where $\bar{\Delta}_n(i)$ is unknown. Then the hypothesis to be tested is that $\bar{\Delta}_n(i) = 0$ for all i. But if we examine the development above, we see that $\{\bar{\Delta}_n(i)\}$ is not completely arbitrary. Instead, $\bar{\Delta}_n(i) = q_n\left(\frac{4i}{K_n-3}\right)$, where q_n (v) is a continuous function of v for $0 < v < 1$. If we have some particular alternative q_n (v) against which to test the hypothesis, a likelihood ratio test can be constructed. If we want to test against a very wide class of alternatives, we could apply one of various nonparametric tests. For example, we could base a test on the total number of runs of positive and negative elements in the sequence $\{\bar{Q}_n(i)\}$. If the hypothesis is true, there should be a relatively large number of runs, but if the hypothesis is false, neighboring $\bar{Q}_n(i)$'s would tend to have the same sign, decreasing the total number of runs. Other tests for an analogous problem are developed in [7].

In the case where $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, all the conditions imposed above hold if we take $p_n = 1 - q_n = 0(n^{-\rho})$, $\epsilon = \frac{1}{2} - \Delta_1$, $\delta = \frac{1}{12} - \frac{\Delta_1}{2} - \Delta_2$, $\gamma = \frac{\Delta_2}{10} - \Delta_3$, $\rho = \frac{\Delta_2}{10} - \Delta_3 - \Delta_4$, where Δ_1, Δ_2, Δ_3, Δ_4 are very small positive values chosen so that $\epsilon > 0$, $\delta > 0$, $\gamma > 0$, and $\rho > 2\Delta_1$.

REFERENCES

[1] Hensler, G.L., K.G. Mehrotra and J.E. Michalek, "A Goodness of Fit Test for Multivariate Normality," Communications in Statistics, *A6,* 33-41 (1977).

[2] Jakobovits, R.H., "Goodness of Fit Tests for Composite Hypotheses Based on an Increasing Number of Order Statistics," Ph.D. Thesis, Cornell University (1977).

[3] Shapiro, S.S. and M.B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika, *52,* 591-611 (1965).

[4] Weiss, L., "Statistical Procedures Based on a Gradually Increasing Number of Order Statistics," Communications in Statistics, *2,* 95-114 (1973).

[5] Weiss, L. "The Asymptotic Sufficiency of a Relatively Small Number of Order Statistics in Tests of Fit," Annals of Statistics, *2,* 795-802 (1974).

[6] Weiss, L., "Testing Fit with Nuisance Location and Scale Parameters," Naval Research Logistics Quarterly, *22,* 55-63 (1975).

[7] Weiss, L., "Asymptotic Properties of Bayes Tests of Nonparametric Hypotheses," *Statistical Decision Theory and Related Topics, II* Academic Press, 439-450 (1977).

[8] Weiss, L., "The Asymptotic Distribution of Order Statistics," Naval Research Logistics Quarterly, *26,* 437-445 (1979).

ON A CLASS OF NASH-SOLVABLE BIMATRIX GAMES AND SOME RELATED NASH SUBSETS

Karen Isaacson and C. B. Millham

Washington State University
Pullman, Washington

ABSTRACT

This work is concerned with a particular class of bimatrix games, the set of equilibrium points of which games possess many of the properties of solutions to zero-sum games, including susceptibility to solution by linear programming. Results in a more general setting are also included. Some of the results are believed to constitute interesting potential additions to elementary courses in game theory.

1. INTRODUCTION

A bimatrix game is defined by an ordered pair $<A,B>$ of $m \times n$ matrices over an ordered field F, together with the Cartesian product $X \times Y$ of all m-dimensional probability vectors $x \in X$ and all n-dimensional probability vectors $y \in Y$. If player 1 chooses a strategy (probability vector) x and player 2 chooses a strategy y, the payoffs to the two players, respectively, are xAy and xBy, where x and y are interpreted appropriately as row or column vectors. A pair $<x^*,y^*>$ in $X \times Y$ is an *equilibrium point* of the game $<A,B>$ if $x^*Ay^* \geqslant xAy^*$ and $x^*By^* \geqslant x^*By$, for all probability vectors x and y.

A Nash-solvable bimatrix game is one in which, if $<x^*,y^*>$ and $<x',y'>$ are both equilibrium points, then so are $<x^*,y'>$ and $<x',y^*>$. It is well known that 0-sum bimatrix games ($a_{ij} + b_{ij} = 0$, all i,j) are Nash-solvable, and that this property extends to constant-sum games ($a_{ij} + b_{ij} = k$, all i,j, for some $k \in F$). It is also well known that in the constant-sum case all equilibrium points are equivalent in that they provide the same payoffs to both players. This work generalizes, slightly, that contained in such sources as Luce and Raiffa (9) and Burger (2), and represents a very small step toward the solution of the open problem of characterizing Nash-solvable games. In the following, $A_{i.}$ will be the ith row of A and $A_{.j}$ the jth column of A, and similarly for B. The inner product of 2 vectors u, v in E^n will be denoted by (u,v). The ordered pair is $<u,v>$.

2. ROW-CONSTANT-SUM BIMATRIX GAMES

DEFINITION 1: An $m \times n$ bimatrix game $<A,B>$ is row-constant-sum if, for each i, $i = 1, \ldots m$, there is a $k_i \in F$ such that $a_{ij} + b_{ij} = k_i$, $j = 1, \ldots n$.

THEOREM 1: Let $<x^*,y^*>$ and $<x',y'>$ be two equilibrium points for a row-constant-sum game $<A,B>$. Then $<x^*,y^*>$ and $<x',y'>$ are interchangeable, and they are equivalent for P1 (player 1). They are equivalent for P2 (player 2) if and only if $\sum_{i=1}^{m} x_i^* k_i = \sum_{i=1}^{m} x_i' k_i$.

PROOF: It is well known and easily proved that $<x^*,y^*>$ is an equilibrium point for $<A,B>$ if and only if $x_i^* > 0$ implies that $(A_{i\cdot}, y^*) = \max_k (A_{k\cdot}, y^*)$ and $y_j^* > 0$ implies that $(x^*,B_{\cdot j}) = \max_k (x^*,B_{\cdot k})$, for all i, j. Accordingly, let $\beta^* = x^*By^*$. Then $y_j^* > 0$ implies $(x^*,B_{\cdot j}) = \beta^* = \sum_i x_i^* k_i - (x^*, A_{\cdot j}) \geqslant \sum_i x_i^* k_i - (x^*,A_{\cdot r})$ for all r, or $(x^*, A_{\cdot r}) \geqslant (x^*, A_{\cdot j})$, and $x_i^* > 0$ implies $(A_{i\cdot},y^*) = \alpha^* = \max_k (A_{k\cdot}, y^*)$. If $<x',y'>$ is any equilibrium point, then we have that $x^*Ay^* \geqslant x'Ay^*$ (because x^* is in equilibrium with y^*) $\geqslant x'Ay'$ (because y' is in equilibrium with x' and by the above argument) $\geqslant x^*Ay'$ (because x' is in equilibrium with y') $\geqslant x^*Ay^*$ (because y^* is in equilibrium with x^* and by the above argument). Thus $<x^*,y^*>$ and $<x',y'>$ are interchangeable for P1, and equivalent for P1. To show they are interchangeable for P2, note that $x'By' = \sum_i x_i' k_i - x'Ay' = \sum_i x_i' k_i - x'Ay^*$, or $x'By' = x'By^*$. One can similarly show that $x^*By^* = x^*By'$, completing this part of the proof.

Suppose now that $\sum_i x_i' k_i = \sum_i x_i^* k_i$. Since $x^*Ay^* = x'Ay^*$, we have that $\sum_i x_i' k_i - x'Ay^* = \sum_i x_i^* k_i - x^*Ay^*$, or $x'By^* = x^* By^*$, and equivalence follows.

On the other hand, suppose $x^*Ay^* = x^*Ay' = x'Ay^* = x'Ay'$, $x^*By^* = x^*By' = x'By^* = x'By'$. Then $\sum_i x_i^* k_i - x^*Ay^* = \sum_i x_i' k_i - x'Ay^*$. Since $x'Ay^* = x^*Ay^*$, it follows that $\sum_i x_i^* k_i = \sum_i x_i' k_i$, and the proof is complete.

It is well known that, if A $(=-B)$ is the payoff matrix for a zero-sum game, optimal strategies $<x^*,y^*>$ for the game satisfy the so-called "saddle-point" property: $x^*Ay \geqslant x^*Ay^* \geqslant xAy^*$ for all probability vectors x and y, and that, conversely, if $<x^*,y^*>$ is a saddle-point of the function xAy, then $<x^*,y^*>$ is a solution to the game A.

THEOREM 2: $<x^*,y^*>$ is an equilibrium point of the row-constant-sum game $<A,B>$ if, and only if, $<x^*,y^*>$ is a saddle-point of the function $\Phi(x,y) = xAy$.

PROOF: If $<x^*,y^*>$ is an equilibrium point of $<A,B>$, then $x^*Ay^* \geqslant xAy^*$ for all $x \in X$, from which half of one implication follows. Now, let K be the $m \times n$ matrix

$$K = \begin{pmatrix} k_1 & k_1 \ldots k_1 \\ k_2 & k_2 \ldots k_2 \\ \ldots\ldots\ldots\ldots \\ k_n & k_n \ldots k_n \end{pmatrix} \quad \text{of row constants } k_i,\ a_{ij} + b_{ij} = k_i.$$

Since $x^*By^* \geqq x^*By$ for all $y \in Y$, we have $x^*(K - A)y^* \geqslant x^*(K - A)y$, from which $x^*Ay \geqslant x^*Ay^*$ $\left(\text{since } x^*Ky^* = x^*Ky = \sum_i x_i^* k_i\right)$. This completes one implication. Suppose now that $<x^*,y^*>$ is a saddle-point of Φ. From $x^*Ay \geqslant x^*Ay^*$ it follows that $y_j^* = 0$ if $(x^*,A_{\cdot j}) > \alpha' = \min_k (x^*,A_{\cdot k})$, from which, if $y_j^* > 0$, $\sum_{i=1}^m x_i^* k_i - (x^*,A_{\cdot j}) \geqslant \sum_i x_i^* k_i - (x^*,A_{\cdot k})$ for all k, or $(x^*,B_{\cdot j}) \geqslant (x^*,B_{\cdot k})$ for all k. Finally, it follows from $x^*Ay^* \geqslant xAy^*$ for all x that $x_i^* = 0$ if $(A_{i\cdot},y^*) < \max_k (A_{k\cdot},y^*)$, and the proof is complete.

The implication is that any solution of A as a 0-sum game is also an equilibrium point of the row-constant-sum bimatrix game $<A,B>$, and conversely. Thus, a solution of A found by

linear programming will provide an equilibrium point $<x^*,y^*>$ for $<A,B>$ and the payoff α for P1. The payoff β for P2 must be calculated via x^*By^*, or via $\sum_i x_i^* k_i - \alpha$.

3. A SOMEWHAT MORE GENERAL SETTING

We now consider the $m \times n$ matrix A, we let B be $m \times n$ (not necessarily in row-constant-sum with A) and we henceforth let $X \times Y$ be the set of solutions to A as a 0-sum game. The following theorem then follows.

THEOREM 3: Let $<x^*,y^*> \in X \times Y$. In order for $<x^*,y^*>$ to be an equilibrium point of $<A,B>$ regarded as a bimatrix game, it is necessary and sufficient that $x^*By^* \geqslant x^*By$ for all probability vectors y, or, for $x^*(-B)y \geqslant x^*(-B)y^*$. It is clearly sufficient for $<x^*,y^*>$ to also be a solution to $(-B)$, regarded as a 0-sum game.

The proof is omitted, as it follows immediately from the definition of equilibrium point. The following comment is made, however: if $<A,B>$ is row-constant-sum, a point $<x^*,y^*>$ that solves A as a 0-sum game and is an equilibrium point of $<A,B>$, will not necessarily solve $(-B)$ as a 0-sum game, because the condition $x^*(-B)y^* \geqslant x(-B)y^*$ holds if and only if $x^*Ay^* - \sum_{i=1}^{m} x_i^* k_i \geqslant xAy^* - \sum_i x_i k_i$, or $x^*Ay^* - xAy^* \geqslant \sum_{i=1}^{m} k_i(x_i^* - x_i)$. Thus, the condition that $<x^*,y^*>$ also solve $(-B)$ as a 0-sum game is extremely strong. This illustrates a major difference between the constant-sum case (in which the above condition will hold if $<x^*,y^*>$ solves A as a 0-sum game) and the row-constant-sum case. It is also logical to ask if there are conditions on A and B which would cause an equilibrium point of $<A,B>$ to also solve A and $-B$ as separate 0-sum games. The conditions are inescapable: $y_i^* > 0$ must imply $(x^*,A_{\cdot j}) = \min_k (x^*,A_{\cdot k})$ and $x_i^* > 0$ must imply $(B_{i\cdot},y^*) = \min_k (B_{k\cdot},y^*)$. Since, for example, to be an equilibrium point of $<A,B>$ it is necessary that $y_j^* > 0$ imply $(x^*,B_{\cdot j}) = \max_k (x^*,B_{\cdot k})$, any game satisfying these conditions must be heavily restricted. Finally, it is noted that if there are common saddle-points of A and $(-B)$, which are therefore equilibrium points of the bimatrix game $<A,B>$, each of these saddle-points will necessarily provide the same payoffs α, β to the respective players (note the contrast of the row-constant-sum case with the constant-sum case).

DEFINITION 2: A Nash Subset for a game $<A,B>$ is a set $S = \{<x,y>\}$ of equilibrium points for $<A,B>$ such that, if $<x,y>$ and $<x',y'>$ are in S, so are $<x,y'>$ and $<x',y>$. See (6) and (13) for related material.

THEOREM 4: Let A and B be $m \times n$ matrices over the ordered field F, and let $X \times Y$ be the set of all solutions to A regarded as a 0-sum game. In order for $X \times Y$ to constitute a Nash subset of equilibrium points for $<A,B>$, regarded as a bimatrix game, it is necessary and sufficient that $K(X) = \{k \mid (x,^*A_{\cdot k}) = \min_i (x^*,A_{\cdot i}),$ all $x^* \in X\} \subset K'(X) = \{k \mid (x^*,B_{\cdot k}) = \max_i (x^*,B_{\cdot i}),$ all $x^* \in X\}$.

PROOF: Write $K = K(X)$, $K' = K'(X)$, and let $K \subset K'$. Then because any $<x^*,y^*>$ in $X \times Y$ solves A as a 0-sum game, $x^*Ay \geqslant x^*Ay^* \geqslant xAy^*$ for all $<x^*,y^*>$ in $X \times Y$ and all probability vectors x, y. Also, $y_j^* = 0$ if $(x^*,A_{\cdot j}) > \min_k (x^*,A_{\cdot k})$, or if $j \notin K \subset K'$. Hence $y_j^* = 0$ if $(x^*,B_{\cdot j}) < \max_i (x^*,B_{\cdot i})$ for all $y^* \in Y$, any $x^* \in X$, and $<x^*,y^*>$ is an equilibrium point for $<A,B>$, for all $<x^*,y^*> \in X \times Y$. Suppose there exists $k' \in K - K'$, so that for

some $x^* \in X, (x^*,B._k) < \max_i (x^*,B._i)$ but $(x^*,A._k) = \min_i (x^*,A._i)$. Since it is known that there exists $y' \in Y$ (see (1), page 52) such that $y'_k > 0$, it follows that y' cannot be in equilibrium with x^* for $<A,B>$ regarded as a bimatrix game, a contradiction. This completes the proof.

COROLLARY 1: Let $X^* \times Y^*$ be any subset of $X \times Y$, the set of all solutions to A regarded as a 0-sum game. In order for $X^* \times Y^*$ to be a set of interchangeable equilibrium points (a Nash subset) for $<A,B>$ regarded as a bimatrix game, it is sufficient that $K(X^*) = \{k \mid (x^*,A._k) = \min_i (x^*,A._i)$ for all $x^* \in X^*\} = K'(X^*) = \{k \mid (x^*,B._k) = \max_i (x^*,B._i)$ for all $x^* \in X^*\}$.

COROLLARY 2: Let $X' \subset X$, and let $K'(X')$ be defined as above, and let $Y' = \{y \in Y \mid y_j > 0$ implies $j \in K'(X')\}$. Then $X' \times Y'$ is a Nash subset for $<A,B>$.

Finally, we consider the construction of all matrices B such that $X \times Y$, the set of solutions to A as a 0-sum game, will also be a set of equilibrium points for $<A,B>$ regarded as a bimatrix game.

THEOREM 5: Let A be an $m \times n$ matrix over F, with $X \times Y$ its solutions as a 0-sum game. Then a matrix B can be constructed such that $X \times Y$ is a Nash subset for $<A,B>$ regarded as a bimatrix game. The equilibrium points $<x,y>$ in $X \times Y$ may or may not be equivalent for P2, depending on construction. Further, all matrices B such that $X \times Y$ is a Nash subset for $<A,B>$ are constructed as described.

PROOF: Let $x^1, x^2, \ldots x^k$ be the extreme points of X, and assume that $x^1, \ldots x^r$, $r \leqslant k$, are a maximal linearly-independent subset of $x^1, \ldots x^k$. Let $\chi = \begin{pmatrix} x^1 \\ \cdot \\ \cdot \\ \cdot \\ x^r \end{pmatrix}$ be regarded as the matrix of a linear transformation from E^m to E^r, taken with respect to a basis of unit vectors, and let $c^1, c^2, \ldots c^{m-r}$ be a basis for the nullspace of χ. Let $\beta_1, \beta_2, \ldots \beta_r$ be scalars. Let $y^1, y^2, \ldots y^l$ be the extreme points of the set Y, and let $K_y j = \{i \mid y^j_i > 0\}$. Let $K_Y = \bigcup_{i=1}^{l} K_y i$. Let $D = \{d \mid (x^i,d) = \beta_j, 1 \leqslant j \leqslant r\}$, and let $d^1, \ldots d^{m-r+1}$ be $m - r + 1$ (if some $\beta_j \neq 0$) linearly-independent solutions to the system of r equations in m variables. For $j \in K_y$, let $B._j = \sum_{i=1}^{m-r+1} \alpha_{ji} d^i + \sum_{i=1}^{m-r} \lambda_{ji} c^i$ where $\sum_{i=1}^{m-r+1} \alpha_{ji} = 1$ $\left(\text{or at least, } \sum_i \alpha_{ji} = \alpha \text{ for some } \alpha \neq 0\right)$, all j. Then, if $x \in X$, there are scalars γ_i, $i = 1, \ldots r$, such that $x = \sum_{i=1}^{r} \gamma_i x^i$, and for $j \in K_Y, (x, B._j) = \left(\sum_{i=1}^{r} \gamma_i x^i, B._j\right) = \left(\sum_{i=1}^{r} \gamma_i x^i, \sum_{p=1}^{m-r+1} \alpha_{jp} d^p + \sum_{i=1}^{m-r} \lambda_{ji} c^i\right) = \sum_{i=1}^{r} \gamma_i \beta_i$ (if $\alpha = 1$). After all $B._j$, $j \in K_Y$, have been constructed, for $j \notin K_Y$, let $B._j$ be such that $(x^i,B._j) \leqslant (x^i,B._h)$, $h \in K_Y$, for all extreme points x^i, $i = 1, \ldots k$. Then, for all $y^* \in Y$, $x^* \in X$ $\left(\text{with } x^* = \sum_{i=1}^{r} \gamma^*_i x^i\right)$, $x^*By^* = \sum_{i=1}^{r} \gamma^*_i \beta_i \geqslant x^*By$ for all probability vectors y. Hence $X \times Y$ is a set of interchangeable equilibrium points for $<A,B>$ that would, for example, be equivalent if $\beta_i = \beta_j$ for all i, j.

Finally, suppose there is a matrix B such that $X \times Y$ is a Nash subset for $<A,B>$ but which does not have the above construction. Then there is a column $B._j$, $j \in K_Y$, such that

either $B_{.j} \neq \sum_{i=1}^{m-r+1} \alpha_{ji} d^i + \sum_{i=1}^{m-r} \lambda_{ji} c^i$ for any coefficients α_{ji}, or $B_j = \sum_{i=1}^{m-r+1} \alpha_{ji} d^i + \sum_{i=1}^{m-r} \lambda_{ji} c^i$ but $\sum_{i=1}^{m-r+1} \alpha_{ji} = \alpha_j \neq \alpha_k = 1$, $k \in K_Y$, $k \neq j$. In the first instance we note $(x^l, B_{.j}) = \beta_l$, $l = 1, \ldots r$, and we contradict the assumption that $d^1, \ldots d^{m-r+1}$ are a maximal linearly-independent set of solutions to $(x^j, d) = \beta_j$, $j = 1, \ldots r$. In the second instance, if $\sum_{i=1}^{m-r+1} \alpha_{ji} = \alpha_j \neq 1$, let $x = \sum_{i=1}^{r} \gamma_i x^i$. Then $(x, B_{.j}) = \alpha_j \sum_{i=1}^{r} \gamma_i \beta_i \neq (x, B_{.k}) = \sum_{i=1}^{r} \gamma_i \beta_i$ for other $k \in K_Y$, so that any equilibrium strategy y will either exclude j, or include j and exclude any k such that $\alpha_k = 1$. Either contradicts the definition of K_Y.

Note that the matrix A is used only to define $X \times Y$. Given the set of $X \times Y$, it follows that both A and B could be constructed as described, assuming the appropriate dimensionality conditions.

4. CONCLUSIONS

It is hoped that this slight extension of previously published material regarding Nash-solvable bimatrix games will lend itself to inclusion in future texts in game theory and operations research covering 2-person, 0-sum finite games (matrix games). Clearly, nearly any statement that can be made about solutions of matrix games can also be made about the somewhat more interesting row-constant-sum bimatrix case, and the usual methods for finding such solutions carry over with the minor modifications indicated. The reader is also referred to the excellent text by Vorobjev (21), and his discussion on "almost antagonistic" bimatrix games (pp. 103-115) for related interesting material.

BIBLIOGRAPHY

[1] Bohnenblust, H.F., S. Karlin, and L.S. Shapley, "Solutions of Discrete, Two-Person Games," *Contributions to the Theory of Games*, Annals of Mathematics, Studies *24*, Princeton University Press (1950).

[2] Burger, E. *Theory of Games.* Prentice-Hall, Englewood Cliffs, New Jersey (1963).

[3] Gale, D. and S. Sherman, "Solutions of Finite Two-Person Games," *Contributions to the Theory of Games,* Annals of Mathematics, Studies *24*, Princeton University Press (1950).

[4] Heuer, G.A., "On Completely Mixed Strategies in Bimatrix Games," The Journal of the London Mathematical Society, *2,* 17-20 (1975).

[5] Heuer, G.A., "Uniqueness of Equilibrium Points in Bimatrix Games," International Journal of Game Theory, *8*, 13-25 (1979).

[6] Heuer, G.A., and C.B. Millham, " On Nash Subsets and Mobility Chains in Bimatrix Games," Naval Research Logistics Quarterly *23,* 311-319 (1976).

[7] Kuhn, H.W., "An Algorithm for Equilibrium Points in Bimatrix Games," Proceedings of the National Academy of Sciences, *47,* 1656-1662 (1961).

[8] Lemke, C.E. and J.T. Howson, Jr., "Equilibrium Points of Bimatrix Games," Journal of the Society for Industrial and Applied Mathematics, *12,* 413-423 (1964).

[9] Luce, R.D., and H. Raiffa, *Games and Decisions*, John Wiley and Sons, New York (1957).

[10] Mangasarian, O.L., "Equilibrium Points of Bimatrix Games," Journal of the Society for Industrial and Applied Mathematics, *12,* 778-780 (1964).

[11] Millham, C.B., "On the Structure of Equilibrium Points in Bimatrix Games," SIAM Review *10,* 447-449 (1968).

[12] Millham, C.B., "Constructing Bimatrix Games with Special Properties," Naval Research Logistics Quarterly *19,* 709-714 (1972).
[13] Millham, C.B., "On Nash Subsets of Bimatrix Games," Naval Research Logistics Quarterly *21,* 307-317 (1974).
[14] Mills, H., "Equilibrium Points in Finite Games," Journal of the Society for Industrial and Applied Mathematics, *8,* 397-402 (1960).
[15] Nash, J.F. Jr., "Two-Person Cooperative Games," Econometrica, *21,* 128-140 (1953).
[16] Owen, G., "Optimal Threat Strategies in Bimatrix Games," International Journal of Game Theory, *1,* 3-9 (1971).
[17] Pugh, G.E. and J.P. Mayberry, "Theory of Measure of Effectiveness for General-Purpose Military Forces, Part 1: A Zero-Sum Payoff Appropriate for Evaluating Combat Strategies," Operations Research *21,* 867-885 (1973).
[18] Raghavan, T.E.S., "Completely Mixed Strategies in Bimatrix Games," The Journal of the London Mathematical Society, *2,* 709-712 (1970).
[19] von Neumann, J. and O. Morganstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, New Jersey (1953), 3rd Ed.
[20] Vorobjev, N.N., "Equilibrium Points in Bimatrix Games," Theoriya Veroyatnostej i ee Primeneniya *3,* 318-331 (1958).
[21] Vorobjev, N.N., *Game Theory* Springer-Verlag, New York, Heidelberg, Berlin (1977).

OPTIMALITY CONDITIONS FOR CONVEX SEMI-INFINITE PROGRAMMING PROBLEMS*

A. Ben-Tal

Department of Computer Science
Technion—Israel Institute of Technology
Haifa, Israel

L. Kerzner

National Defence
Ottawa, Canada

S. Zlobec

Department of Mathematics
McGill University
Montreal, Quebec, Canada

ABSTRACT

This paper gives characterizations of optimal solutions for convex semi-infinite programming problems. These characterizations are free of a constraint qualification assumption. Thus they overcome the deficiencies of the semi-infinite versions of the Fritz John and the Kuhn-Tucker theories, which give only necessary or sufficient conditions for optimality, but not both.

1. INTRODUCTION

A mathematical programming problem with infinitely many constraints is termed a "semi-infinite programming problem." Such problems occur in many situations including production scheduling [10], air pollution problems [6],[7], approximation theory [5], statistics and probability [9]. For a rather extensive bibliography on semi-infinite programming the reader is referred to [8].

The purpose of this paper is to give necessary and sufficient conditions of optimality for convex semi-infinite programming problems. It is well known that the semi-infinite versions of both the Fritz John and the Kuhn-Tucker theories fail to characterize optimality (even in the linear case) unless a certain hypothesis, known as a "constraint qualification," is imposed on the problem, e.g. [4],[12]. This paper gives a characterization of optimality without assuming a constraint qualification.

*This research was partially supported by Project No. NR047-021, ONR Contract N00014-75-C0569 with the Center for Cybernetic Studies, The University of Texas and by the National Research Council of Canada.

Characterization theorems without a constraint qualification for ordinary (i.e. with a finite number of constraints) mathematical programming problems have been obtained in [1]. It should be noted that the analysis of the semi-infinite case is significantly different; the special feature being here the topological properties of all constraint functions including the particular role played by the *nonbinding* constraints.

The optimality conditions are given in Section 2 for differentiable convex semi-infinite programming programs, whose constraint functions have the "uniform mean value property." This class of programs is quite large and it includes programs with arbitrary convex objective functions and linear or strictly convex constraints. For a particular class of such programs, namely the programs with "uniformly decreasing" constraint functions, the optimality conditions can be strengthened, as shown in Section 4. A comparison with the semi-infinite analogs of the Fritz John and Kuhn-Tucker theories is presented in Section 5. An application to the problem of best linear Chebyshev approximation with constraints is demonstrated in Section 6. A linear semi-infinite problem taken from [4], for which the Kuhn-Tucker theory fails, is solved in this section using our results.

2. OPTIMALITY CONDITIONS FOR PROGRAMS HAVING UNIFORM MEAN VALUE PROPERTY

Consider the convex semi-infinite programming problem

(P)

$$\text{Min } f^o(x)$$

s.t.

$$f^k(x,t) \leqslant 0 \text{ for all } t \in T_k,\ k \in P \triangleq \{1, \ldots, p\}$$
$$x \in R^n$$

where

f^o is convex and differentiable,

$f^k(x,t)$ is convex and differentiable in x for every $t \in T_k$ and continuous in t for every x,

T_k is a compact subset of R^l $(l \geqslant 1)$.

The feasible set of problem (P) is

$$F = \{x \in R^n : f^k(x,t) \leqslant 0 \text{ for all } t \in T_k,\ k \in P\}.$$

Note that F is a convex set being the intersection of convex sets.

For $x^* \in F$,

$$T_k^* \triangleq \{t \in T_k : f^k(x^*,t) = 0\},$$
$$P^* \triangleq \{k \in P : T_k^* \neq 0\}.$$

A vector $d \in R^n$ is called a *feasible direction at* x^* if $x^* + d \in F$. For a given function $f^k(\cdot,t)$, $k \in \{0\} \cup P$ and for a fixed $t \in T_k$, we define

$$D_k(x^*,t) \triangleq \{d \in R^n : \exists\, \bar{\alpha} > 0 \ni f^k(x^* + \alpha d,t) = f^k(x^*,t) \text{ for all } 0 \leqslant \alpha \leqslant \bar{\alpha}\}.$$

This set is called the *cone of directions of constancy* in [1], where it has been shown that, for a differentiable convex function $f^k(\cdot,t)$, it is a convex cone contained in the subspace

$$\{d: d^t \nabla f^k(x^*,t) = 0\}.$$

Furthermore, if $f^k(\cdot,t)$ is an analytic convex function, then $D_k(x^*,t)$ is a subspace (not depending on x^*), see [1, Example 4]. In the sequel the derivative of f with respect to x, i.e. $\nabla_x f(x,t)$, is denoted by $\nabla f(x,t)$.

Optimality conditions will be given for problem (P) if the constraint functions have the "uniform mean value property" which is defined as follows.

DEFINITION 1: Let T be a compact set in R^l. A function $f: R^n \times T \to R$ has the uniform mean value property at $x \in R^n$ if, for every nonzero $d \in R^n$ and every $\alpha > 0$, there exists $\hat{\alpha} = \hat{\alpha}(d,\alpha)$, $0 < \hat{\alpha} \leqslant \alpha$ such that

$$\text{(MV)} \qquad \frac{f(x+\alpha d,t) - f(x,t)}{\alpha} \geqslant d^t \nabla f(x+\hat{\alpha} d,t) \text{ for every } t \in T.$$

If $f(\cdot,t)$ is a linear function in x for every $t \in T$, i.e. if f is of the form

$$f(x,t) = g(t) + \sum_{i=1}^{n} x_i g_i(t),$$

or if $f(\cdot,t)$ is a differentiable strictly convex function in x for every $t \in T$, i.e. if

$$f(\lambda x + (1-\lambda)y,t) < \lambda f(x,t) + (1-\lambda) f(y,t) \text{ for every } t \in T$$

where $y \in R^n$ is arbitrary, $y \neq x$, $0 < \lambda < 1$, and if $f(x,\cdot)$ is continuous in t for every x, then f has the uniform mean value property. For a linear function f, one finds $d^t \nabla f(x+\alpha d,t) = \sum_{i=1}^{n} d_i g_i(t)$ and (MV) is obviously satisfied. The mean value property for strictly convex functions follows immediately from e.g. [14, Corollary 25.5.1 and Theorem 25.7].

EXAMPLE 1: Function

$$f^1(x,t) = t^2[(x-t)^2 - t^2] \text{ for every } t \in T = [0,1]$$

is neither linear nor strictly convex in $x \in R$ for every $t \in T$. However f^1 has the uniform mean value property. Function

$$f^2(x_1,x_2,t) = \begin{cases} x_1^2 + t x_2(x_2 - t) \text{ if } x_2 < \frac{1}{2}t \\ x_1^2 + \dfrac{t^3}{(2-t)^2}(x_2 - t + 1)(x_2 - 1) \text{ if } x_2 \geqslant \frac{1}{2}t \end{cases}$$

for every $t \in T = [0,1]$ does not have the uniform mean value property at the origin. Note that f^2 is convex and differentiable in $x \in R^2$ for every $t \in T$ and continuous in $t \in T$ for every x. This function has provided counterexamples to some of our early conjectures.

Optimality conditions will now be given for problem (P).

THEOREM 1: *Let* x^* be a feasible solution of problem (P) where f^k, $k \in P^*$ have the uniform mean value property. Then x^* is an optimal solution of (P) if, and only if, for every $\alpha^* > 0$ the system

(A) $d^t \nabla f^o(x^*) < 0$,

(B) $d^t \nabla f^k(x^* + \alpha^* d, t) \leqslant 0$ for all $t \in T_k^*$,

(C) $$\frac{d^t \nabla f^k(x^* + \alpha^* d, t)}{f^k(x^*, t)} \geqslant -\frac{1}{\alpha^*} \quad \text{for all } t \in T_k \backslash T_k^*, \quad k \in P^*$$

is inconsistent.

PROOF: We will show that x^* is nonoptimal if, and only if, there exists $\alpha^* > 0$ such that the system (A), (B), (C) is consistent. A feasible x^* is nonoptimal if, and only if, there exist $\bar{\alpha} > 0$ and $d \in R^n$, $d \neq 0$, such that

(1) $f^o(x^* + \bar{\alpha} d) < f^o(x^*)$

(2) $f^k(x^* + \bar{\alpha} d, t) \leqslant 0$ for every $t \in T_k$, $k \in P$.

By the convexity of f^o and the gradient inequality, the existence of $\bar{\alpha} > 0$ satisfying (1) is equivalent to

$$d^t \nabla f^o(x^*) < 0.$$

By the continuity of $f^k(\cdot, t)$, $k \in P$, the constraints with $k \in P \backslash P^*$ can be omitted from discussion. We consider (2), for some given $k \in P^*$, and discuss separately the two cases: $t \in T_k^*$ and $t \in T_k \backslash T_k^*$. Thus (2) can be written

(2-a) $f^k(x^* + \bar{\alpha} d, t) \leqslant 0$ for every $t \in T_k^*$

(2-b) $f^k(x^* + \bar{\alpha} d, t) \leqslant 0$ for every $t \in T_k \backslash T_k^*$.

Consider first (2-a) for some fixed $k \in P^*$. By the convexity and uniform mean value property of f^k,

(3) $f^k(x^* + \bar{\alpha} d, t) \geqslant f^k(x^*, t) + \bar{\alpha} d^t \nabla f^k(x^* + \alpha_k d, t)$ for all $t \in T_k^*$

and for some

$$0 < \alpha_k \leqslant \bar{\alpha}.$$

Since $t \in T_k^*$ and $\bar{\alpha} > 0$, (2-a) implies

(4) $d^t \nabla f^k(x^* + \alpha_k d, t) \leqslant 0$.

Denote

(5) $$\hat{\alpha} = \min_{k \in P^*} \{\alpha_k\}.$$

Clearly, $\hat{\alpha}$ always exists (since P is finite) and it is positive. By the convexity of $f^k(\cdot, t)$, (5) and (4),

(6) $d^t \nabla f^k(x^* + \hat{\alpha} d, t) \leqslant d^t \nabla f^k(x^* + \alpha_k d, t) \leqslant 0$.

On the other hand, the existence of $\alpha^* > 0$ such that, for some $t \in T_k^*$ and all $k \in P^*$,

$$d^t \nabla f^k(x^* + \alpha^* d, t) \leqslant 0$$

implies (2-a) with $0 < \bar{\alpha} \leqslant \alpha^*$.

It is left to show that the existence of $\bar{\alpha} > 0$, such that (2-b) holds, is equivalent to the existence of $\tilde{\alpha} > 0$, such that (C) holds. Suppose that (2-b) holds for some $\bar{\alpha} > 0$. Then, by the convexity and uniform mean value property, for $k \in P^*$,

$$0 \geqslant f^k(x^* + \bar{\alpha} d, t) \geqslant f^k(x^*, t) + \bar{\alpha} d' \nabla f^k(x^* + \tilde{\alpha}_k d, t) \text{ for all } t \in T_k \backslash T_k^*$$

and for some

$$0 < \tilde{\alpha}_k \leqslant \bar{\alpha}. \tag{7}$$

Hence,

$$\frac{d' \nabla f^k(x^* + \tilde{\alpha}_k d, t)}{f^k(x^*, t)} \geqslant -\frac{1}{\bar{\alpha}}, \text{ since } t \in T_k \backslash T_k^*$$

$$\geqslant -\frac{1}{\tilde{\alpha}_k}, \text{ by (7).} \tag{8}$$

Denote

$$\tilde{\alpha} = \min_{k \in P^*} \{\tilde{\alpha}_k\} > 0. \tag{9}$$

Using the monotonicity of the gradient of the convex function $f^k(\cdot, t)$, one obtains here

$$\frac{d' \nabla f^k(x^* + \alpha d, t)}{f^k(x^*, t)} \geqslant \frac{d' \nabla f^k(x^* + \tilde{\alpha}_k d, t)}{f^k(x^*, t)} \text{ for every } 0 \leqslant \alpha \leqslant \tilde{\alpha}_k. \tag{10}$$

This gives

$$\frac{d' \nabla f^k(x^* + \tilde{\alpha} d, t)}{f^k(x^*, t)} \geqslant -\frac{1}{\tilde{\alpha}_k}, \text{ by (10) and (8)}$$

$$\geqslant -\frac{1}{\tilde{\alpha}}, \text{ by (9)}$$

which is (C) with $\alpha^* = \tilde{\alpha}$.

Suppose now that (C) is true for some $\alpha^* > 0$. Using again the monotonicity of the gradient of the convex function $f^k(\cdot, t)$, and the fact that $f^k(x^*, t) < 0$ for $t \in T_k \backslash T_k^*$, one easily obtains

$$f^k(x^*, t) + \alpha^* d' \nabla f^k(x^* + \alpha d, t) \leqslant 0, \text{ for every } 0 < \alpha < \alpha^*. \tag{11}$$

But

$$f^k(x^* + \alpha^* d, t) = f^k(x^*, t) + \alpha^* d' \nabla f^k(x^* + \alpha_k d, t),$$

for some particular $0 < \alpha_k < \alpha^*$, $\alpha_k = \alpha_k(t)$

by the mean value theorem

$$\leqslant 0, \text{ by (11)}$$

which is (2-b) with $\bar{\alpha} = \alpha^*$.

Summarizing the above results one derives the following conclusion: If x^* is not optimal then there exists $\alpha^* = \min\{\hat{\alpha}, \tilde{\alpha}\} > 0$ such that the system (A), (B) and (C) is consistent. If there exists $\alpha^* > 0$ such that the system (A), (B) and (C) is consistent, then there exist $\alpha_o > 0$ and $\bar{\alpha} > 0$ such that

$$(12)\qquad \begin{cases} f^0(x^* + \alpha_0 d) < f^0(x^*) \\ f^k(x^* + \bar{\alpha} d, t) \leqslant 0 \text{ for every } t \in T_k^* \\ f^k(x^* + \bar{\alpha} d, t) \leqslant 0 \text{ for every } t \in T_k \backslash T_k^* \\ \qquad\qquad\qquad\qquad\qquad\qquad k \in P^*. \end{cases}$$

If one denotes

$$\hat{\alpha} = \min\{\alpha_0, \bar{\alpha}\} > 0$$

then, again by the convexity of $f^k(\cdot, t)$, $k \in \{0\} \cup P$, (12) can be written

$$\begin{cases} f^0(x^* + \hat{\alpha} d) < f^0(x^*) \\ f^k(x^* + \hat{\alpha} d, t) \leqslant 0 \text{ for every } t \in T_k. \\ \qquad\qquad\qquad\qquad k \in P^* \end{cases}$$

implying that x^* is not optimal.

□

REMARK 1: Since $\nabla f^k(x, \cdot)$ is continuous for every x in some neighbourhood of x^* (this follows from e.g. [14, Theorem 25.7]), condition (C) in Theorem 1 needs checking only at the points $t \in T_k$ which are in

$$N_k \triangleq \bigcup_{t^* \in T_k^*} N(t^*),$$

where $N(t^*)$ is a fixed open neighbourhood of t^*. For the points t in $T_k \backslash N_k$ one can always find α^* which satisfies (C). This follows from the fact that for every $\bar{\alpha}$,

$$(13)\qquad \frac{d' \nabla f^k(x^* + \bar{\alpha} d, t)}{f^k(x^*, t)} \geqslant -M$$

for some positive constant M, by the compactness of $T_k \backslash N_k$. Choose M in (13) large enough, so that

$$(14)\qquad \alpha^* \triangleq \frac{1}{M} \leqslant \bar{\alpha}.$$

Now,

$$\frac{d' \nabla f^k(x^* + \alpha^* d, t)}{f^k(x^*, t)} \geqslant \frac{d' \nabla f^k(x^* + \bar{\alpha} d, t)}{f^k(x^*, t)}, \text{ by (10) and (14)}$$

$$\geqslant -\frac{1}{\alpha^*}, \text{ by (13) and (14).}$$

EXAMPLE 2: The purpose of this example is to show that Theorem 1 fails if the constraint functions do not have the uniform mean value property. Consider

$$\text{Min } -x_2$$

subject to

$$f(x_1, x_2, t) \leqslant 0 \text{ for all } t \in [0, 1]$$

where

$$f(x_1,x_2,t) = \begin{cases} x_1^2 + tx_2(x_2 - t) & \text{if } x_2 < \frac{1}{2}t \\ x_1^2 + \dfrac{t^3}{(2-t)^2}(x_2 - t + 1)(x_2 - 1) & \text{if } x_2 \geqslant \frac{1}{2}t. \end{cases}$$

Function f satisfies the assumptions of problem (P) but it does not enjoy the uniform mean value property. The feasible set is

$$F = \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} : 0 \leqslant x_2 \leqslant 1 \right\}$$

and the optimal solution is $\bar{x} = (0,1)^t$. However, for every $\alpha^* > 0$, the system (A), (B) and (C) is inconsistent at $x^* = 0$, a nonoptimal point. Since $T^* = [0,1]$, condition (C) is here redundant, while (A) and (B) become, respectively,

$$-d_2 < 0$$

$$2\alpha^* d_1^2 + t(2\alpha^* d_2 - t)d_2 \leqslant 0 \text{ if } 2\alpha^* d_2 < t$$

$$2\alpha^* d_1^2 + \frac{t^3}{(2-t)^2}(2\alpha^* d_2 - t)\, d_2 \leqslant 0 \text{ if } 2\alpha^* d_2 \geqslant t.$$

The above system cannot be consistent for some $\alpha^* > 0$, because, if it were, the last inequality would be absurd for small $t \in [0,1]$.

When the constraint functions (but not necessarily the objective function) are linear, i.e. when (P) is of the form

(L)

$$\text{Min } f^o(x)$$

s.t.

$$g_o^k(t) + \sum_{i=1}^{n} x_i g_i^k(t) \leqslant 0, \text{ for all } t \in T_k,\ k \in P$$

then Theorem 1 can be considerably simplified.

COROLLARY 1: Let x^* be a feasible solution of problem (L). Then x^* is optimal if, and only if, the system

(A) $$d^t \nabla f^o(x^*) < 0$$

(B_2) $$\sum_{i=1}^{n} d_i g_i^k(t) \leqslant 0, \text{ for all } t \in T_k^*$$

(C_1) $$\frac{\sum_{i=1}^{n} d_i g_i^k(t)}{g_o^k(t) + \sum_{i=1}^{n} x_i^* g_i^k(t)} \geqslant -1, \text{ for all } t \in T_k \backslash T_k^*,$$

$$k \in P^*$$

is inconsistent.

PROOF: Recall that linear functions have the uniform mean value property. If $f^k(\cdot,t)$ is linear, then for every $t \in T_k$

$$D_k(x,t) = \{d \in R^n: d' \nabla f^k(x,t) = 0\}.$$

Thus (B) reduces to (B_2). The left hand side of (C) reduces to the left hand side of (C_1), which does not depend on α^*. Moreover, α^* on the right hand side of (C) can be taken $\alpha^* = 1$, because whenever $\bar{d}$ satisfies (A) and (B_2), so does $d = \frac{1}{\alpha^*}\bar{d}$.

□

In many practical situations the sets T_k, $k \in P$ are compact intervals and the sets T_k^*, $k \in P^*$ are finite. (This is always the case when $f(x^*,\cdot)$ are analytic functions not identically zero.) For such cases condition (B) can be replaced by a finite number of linear inequalities.

COROLLARY 2: Let x^* be a feasible solution of problem (P), where f^k, $k \in P^*$ have the uniform mean value property. Suppose that all the sets T_k^*, $k \in P^*$ are finite. Then a feasible solution x^* of problem (P) is optimal if, and only if, for every $\alpha^* > 0$ and for every subset Ω_k of T_k^* the system

(A) $$d' \nabla f^o(x^*) < 0$$

(B_3) $$\begin{cases} d' \nabla f^k(x^*,t) < 0, & t \in \Omega_k \\ d \in D_k(x^*,t), & t \in T_k^* \backslash \Omega_k \end{cases}$$

(C) $$\begin{cases} \dfrac{d' \nabla f^k(x^* + \alpha^* d, t)}{f^k(x^*,t)} \geqslant -\dfrac{1}{\alpha^*} \\ \qquad \text{for all } t \in T_k \backslash T_k^*, \end{cases} \qquad k \in P^*$$

is inconsistent.

An important special case of Corollary 2 is when the sets T_k themselves are finite. Then problem (P) can be reduced to a mathematical program of the form

(MP)

$$\text{Min } f^o(x)$$

s.t.

$$f^k(x) \leqslant 0, \quad k \in P.$$

This is obtained by setting $T_k = \{k_1, k_2, \ldots, k_{\text{card}T_k}\}$ and identifying $\{f^k(x,k_i): k_i \in T_k, k = 1,2,\ldots,p\}$ with $\left\{f^k(x): k \in P \triangleq \{1,2,\ldots,\sum_{k=1}^{p} \text{card } T_k\}\right\}$. Here $P^* = \{k \in P: f^k(x^*) = 0\}$. Also $\{D_k(x^*,k_i): k_i \in T_k, k = 1,2,\ldots,p\}$ is denoted by $\{D_k(x^*): k \in P\}$.

The major difference between the semi-infinite problem (P) and the mathematical problem (MP) is that for the latter the condition (C) is redundant; Theorem 1 then reduces to the following result obtained in [1, Theorem 1].

COROLLARY 3: Consider problem (MP), where $\{f^k: k \in \{0\} \cup P\}$ are differentiable convex functions: $R^n \to R$. A feasible solution x^* of (MP) is optimal if, and only if, for every subset Ω of P^* the system

$$d' \nabla f^o(x^*) < 0$$

$$d' \nabla f^k(x^*) < 0, \; k \in \Omega$$

$$d \in D_k(x^*), \; k \in P^* \backslash \Omega$$

is inconsistent.

PROOF: Here condition (C) becomes

$$\frac{d' \nabla f^k(x^* + \alpha^* d)}{f^k(x^*)} \geqslant -\frac{1}{\alpha^*}, \; k \in P \backslash P^*$$

for some $\alpha^* > 0$. Since here the set $P \backslash P^*$ is finite, and hence compact, the redundancy of condition (C) is shown as in Remark 1.

□

The following result gives a characterization of a unique optimal solution of problem (P).

THEOREM 2: Let x^* be a feasible solution of problem (P), where f^k, $k \in P^*$ have the uniform mean value property. Then x^* is a unique optimal solution of problem (P) if, and only if, for every $\alpha^* > 0$ there is no d satisfying conditions (B), (C) and

(A_1) $\qquad d' \nabla f^o(x^*) < 0$ or $d \in D_o(x^*)$.

PROOF: Suppose that the system (A_1), (B), (C) is inconsistent. Then so is the system (A), (B), (C). Hence, by Theorem 1, x^* is an optimal solution. Suppose that x^* is not a unique optimal solution. Then there exist $\bar{\alpha} > 0$ and $\bar{d} \neq 0$ such that $\bar{x} = x^* + \bar{\alpha}\bar{d}$ is feasible, which implies that $\bar{d}$ satisfies (B), (C) and $f^o(x^*) = f^o(x^* + \bar{\alpha}\bar{d})$. Since the set of all optimal solutions of a convex program is convex, the latter implies $f^o(x^*) = f^o(x^* + \alpha\bar{d})$ for all $0 \leqslant \alpha \leqslant \bar{\alpha}$, i.e., $\bar{d} \in D_o(x^*)$. Thus $\bar{d}$ satisfies (A_1), (B) and (C), which is impossible. Therefore x^* is the unique optimum. The necessity follows by a similar argument.

□

3. OPTIMALITY CONDITIONS FOR STRICTLY CONVEX FUNCTIONS IN THEIR ACTUAL VARIABLES

This section can be skipped without hindering the study of Section 4.

In order to state our next result, which is a characterization of optimality for a subclass of convex functions, i.e. strictly convex functions in their "actual variables", we adopt some notions from [1].

For every $k \in P$ and $t \in T_k$, denote by $[k](t)$ (read "block k"), the following index subset of P: $j \in [k](t)$ if, and only if, $\gamma^k: R \to R$, defined by

$$\gamma^k(\cdot) \triangleq (x_1, \ldots, x_{j-1}, \cdot, x_{j+1}, \ldots, x_n)$$

is not a constant function for some fixed $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$. Thus, for a given $t \in T_k$, $[k](t)$ is the set of indices of those variables on which $f^k(\cdot, t)$ actually depends. These "actual

variables" determine the vector $x_{[k](t)}$, obtained from $x = (x_1, \ldots, x_n)'$ by deleting the variables $\{x_j: j \notin [k](t)\}$, without changing the order of the remaining ones. Similarly, we denote by $f^{[k](t)}: R^{\text{card}[k]} \to R$ the restriction of f^k to $R^{\text{card}[k]}$.

DEFINITION 2: A function $f^k: R^n \times T_k \to R$ is strictly convex in its actual variables if for every $t \in T_k$ its restriction $f^{[k](t)}(\cdot,t)$ is strictly convex.

The above concept will be illustrated by an example.

EXAMPLE 3: Consider

$$f^1(x,t) = x_1^2 + tx_2^2, \quad t \in T = [0,1].$$

Note that function $f^1(\cdot,t)$ is not strictly convex for every $t \in T$. Here

$$[1](t) = \begin{cases} \{1\} \text{ if } t = 0 \\ \{1,2\} \text{ if } t \in (0,1], \end{cases}$$

$$x_{[1](t)} = \begin{cases} (x_1) \text{ if } t = 0 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \text{ if } t \in (0,1] \end{cases}$$

and

$$f^{[1](t)} = \begin{cases} x_1^2 \text{ if } t = 0 \\ x_1^2 + tx_2^2 \text{ if } t \in (0,1], \end{cases}$$

clearly a strictly convex function in its actual variables for every $t \in T$. Hence, f^1 is a strictly convex function in its actual variables.

COROLLARY 4: Let x^* be a feasible solution of problem (P), where $f^k(\cdot,t)$, $k \in P^*$ are strictly convex in their actual variables and have the uniform mean value property. Then x^* is an optimal solution of (P) if, and only if, for every $\alpha^* > 0$ and every subset $\Omega_k \subset T_k^*$ the system

(A) $\quad d' \nabla f^o(x^*) < 0$

(B, Ω) $\quad d' \nabla f^k(x^* + \alpha^* d, t) < 0$ for all $t \in T_k^* \backslash \Omega_k$

(C) $\quad \dfrac{d' \nabla f^k(x^* + \alpha^* d, t)}{f^k(x^*, t)} \geqslant -\dfrac{1}{\alpha^*}$ for all $t \in T_k \backslash T_k^*$

(D, Ω) $\quad d_{[k](t)} = 0$ for all $t \in \Omega_k$, $\quad k \in P^*$

is inconsistent.

PROOF: We know, by Theorem 1, that x^* is nonoptimal if, and only if, there exists $\alpha^* > 0$ such that the system (A), (B), (C) is consistent. In order to prove Corollary 4, it is enough to show that (B) is consistent if, and only if, for some subsets $\Omega_k \subset T_k^*$, $k \in P^*$, the system (B, Ω), (D, Ω) is consistent. Suppose that (B) holds. For every $k \in P^*$ define

$$\hat{\Omega}_k \triangleq \{t \in T_k^*: d' \nabla f^k(x^* + \alpha d, t) = 0 \text{ for all } 0 < \alpha \leqslant \alpha^*\}.$$

Hence, by the mean value theorem, for every $t \in \hat{\Omega}_k$

$$f^k(x^* + \alpha d, t) = f^k(x^*, t) \text{ for all } 0 < \alpha \leqslant \alpha^*.$$

Since $f^k(\cdot,t)$ is strictly convex in its actual variables, this is equivalent to

$$d_{[k](t)} = 0 \text{ for all } t \in \hat{\Omega}_k.$$

If $t \in T_k^* \backslash \hat{\Omega}_k$, then obviously $d' \nabla f^k(x^* + \bar{\alpha} d, t) < 0$ for some $0 < \bar{\alpha} < \alpha^*$, by (B). Thus (B, Ω), (D, Ω) holds for $\Omega_k = \hat{\Omega}_k$. (Note that some or all $\hat{\Omega}_k$'s may be empty.) The reverse statement follows from the observation that $d_{[k](t)} = 0$ implies $d' \nabla f^k(x^* + \alpha^* d, t) = 0$.

□

If a function $f^k(\cdot,t)$ is strictly convex (in all variables $x_1, \ldots, x_n$) for every $t \in T_k$, $k \in P^*$, then $D_k(x^*,t) = \{0\}$. This implies that the system (A), (B, Ω), (C), (D, Ω) is inconsistent for every nonempty Ω_k, $k \in P^*$. Thus condition (D, Ω) is redundant. In fact, condition (C) is also redundant, which follows by the following lemma.

LEMMA 1: Let $f(x,t)$ be convex and differentiable in $x \in R^n$ for every t in a compact set $T \subset R^l$ and continuous in t for every x. If for some $d \in R^n$,

$$d' \nabla f(x^*,t) < 0 \text{ for all } t \in T^* = \{t: f(x^*,t) = 0\}, \tag{15}$$

then there exists $\alpha > 0$ such that

$$f(x^* + \alpha d, t) \leqslant 0 \text{ for all } t \in T \backslash T^*. \tag{16}$$

PROOF: It is enough to show that the hypothesis (15) and the negation of the conclusion (16), which is

"For every $\alpha > 0$ there is $t = t(\alpha) \in T \backslash T^*$ such that $f(x^* + \alpha d, t(\alpha)) > 0$,"

are not simultaneously satisfied. If this were true one would have the following situation:

For every α_n of the sequence $\alpha_n = 2^{-n}$ there is a $t_n = t_n(\alpha_n) \in T \backslash T^*$ such that

$$f(x^* + \alpha_n d, t_n(\alpha_n)) > 0, \quad n = 0, 1, 2, \ldots \tag{17}$$

Since T is compact, $\{t_n\}$ has an accumulation point $\hat{t} \in T$, i.e. there is a convergent subsequence $\{t_{n_i}\}$ with $\hat{t}$ as its limit point. We discuss separately two possibilities and arrive at contradictions in each case.

CASE I: $\hat{t} \in T^*$. Since $f(x^*,\hat{t}) = 0$ and $d' \nabla f(x^*,\hat{t}) < 0$, by (15), there exists $\hat{\alpha} > 0$ such that

$$f(x^* + \hat{\alpha} d, \hat{t}) < 0. \tag{18}$$

For all large values of index i, $\alpha_{n_i} < \hat{\alpha}$ and

$$f(x^*, t_{n_i}) < 0, \tag{19}$$

since $t_{n_i} \in T \backslash T^*$. This implies

$$f(x^* + \hat{\alpha} d, t_{n_i}) > 0. \tag{20}$$

(If (20) were not true, one would have, for some particular n_i,

$$f(x^* + \hat{\alpha} d, t_{n_i}) \leqslant 0. \tag{21}$$

Now $\alpha_{n_i} < \hat{\alpha}$, (19), (21) and the convexity of f imply

$$f(x^* + \alpha_{n_i} d, t_{n_i}) \leqslant 0$$

which contradicts (17).) But (18) and (20) contradict the continuity of $f(x^* + \hat{\alpha} d, \cdot)$.

CASE II: $\hat{t} \in T \backslash T^*$. Since $f(x^*, \hat{t}) < 0$, there exists $\hat{\alpha} > 0$ such that (18) holds, by the continuity of $f(\cdot, \hat{t})$. The rest of the proof is the same as in Case I.

□

A characterization of optimality for strictly convex constraints follows.

COROLLARY 5: Let x^* be a feasible solution of problem (P), where $f^k(\cdot, t)$ are strictly convex for every $t \in T_k$, $k \in P^*$. Then x^* is an optimal solution of (P) if, and only if, for every $\alpha^* > 0$ the system

$$(A) \qquad d' \nabla f^o(x^*) < 0$$

$$(B_1) \qquad d' \nabla f^k(x^*, t) < 0 \text{ for all } t \in T_k^* \quad k \in P^*$$

is inconsistent.

PROOF: First we recall that f^k, $k \in P^*$, under the assumption of the corollary, have the uniform mean value property. If x^* is not optimal, then the system (A), (B_1), (C) is consistent, by Corollary 4. This implies that the less restrictive system (A), (B_1) is consistent. Suppose that the system (A), (B_1) is consistent. Then for every $k \in P^*$ there is $\alpha_k > 0$ such that

$$f^k(x^* + \alpha_k d, t) \leqslant 0 \text{ for all } t \in T_k \backslash T_k^*$$

by Lemma 1. Let

$$\alpha^* \triangleq \min\{\alpha_k : k \in P^*\}.$$

By the convexity of f^k, it follows that

$$f^k(x^* + \alpha^* d, t) \leqslant 0 \text{ for all } t \in T_k \backslash T_k^* \text{ and } k \in P^*.$$

This is equivalent to (C) of Theorem 1 (see (2-b)). Therefore the system (A), (B_1), (C) is consistent. This implies that the system (A), (B), (C) is consistent. (The reader may verify this statement by the technique used in the proof of Lemma 2.) Hence x^* is optimal, by Corollary 4.

□

REMARK 2: Differentiable strictly convex (in all variables!) functions f^k do have the uniform mean value property. However, this is not necessarily true in the case of convex functions with strictly convex restrictions. In particular, function

$$f(x_1, x_2, t) = \begin{cases} x_1^2 + t x_2 (x_2 - t) & \text{if } x_2 < \frac{1}{2} t \\ x_1^2 + \dfrac{t^3}{(2-t)^2} (x_2 - t + 1)(x_2 - 1) & \text{if } x_2 \geqslant \frac{1}{2} t \end{cases}$$

is differentiable and has strictly convex restrictions for every $t \in [0, 1]$. Note that

$$[k](t) = \begin{cases} \{1\} & \text{if } t = 0 \\ \{1, 2\} & \text{if } t \in (0, 1]. \end{cases}$$

But function f does not have the uniform mean value property. One can show, however, that a differentiable function which is strictly convex in its actual variables and such that $[k](t)$ is constant over all compact set T, does have the mean value property.

4. PROGRAMS WITH UNIFORMLY DECREASING CONSTRAINTS

The applicability of Theorem 1 is, in general, obscured by the appearance of parameter α^* in conditions (B) and (C). The purpose of this section is to point out some of the topological difficulties which arise in the removing of α^* from condition (B). A class of convex functions for which the optimality conditions can be stated without reference to α^* in condition (B) will be called the uniformly decreasing functions.

In what follows we assume that $f: R^n \times T \to R$ is convex and differentiable in $x \in R^n$ for every t of a compact set T in R^m. Further, $\nabla f(x^*,t)$ denotes $\nabla f_x(x^*,t)$.

DEFINITION 3: Let $f: R^n \times T \to R$ and $x^* \in R^n$ be such that $T^* \neq \emptyset$. Then for a given $d \in R^n$, $d \neq 0$, the function f is uniformly decreasing at x^* in the direction d, if (i) the set

$$S(x^*,d) \triangleq \{t \in T^*: d' \nabla f(x^*,t) < 0\}$$

is compact and if (ii) there exists $\bar{\alpha} > 0$ such that $f(x^* + \bar{\alpha} d, t) = 0$ for all $t \in T^*$ for which $d \in D(x^*,t)$.

It is not easy to recognize whether a general convex function f is uniformly decreasing.

EXAMPLE 4: Consider the following functions from $R \times R$ into R:

$$f^1(x,t) = t^2[(x-t)^2 - t^2], \quad t \in T \text{ (used in Example 1)}$$

$$f^2(x,t) = x^2 - tx, \qquad t \in T$$

$$f^3(x,t) = -tx, \qquad t \in T.$$

These functions are all convex, f^2 is actually strictly convex and f^3 linear in x for every $t \in T$. If $T = [0,1]$, then neither function is uniformly decreasing at $x^* = 0$ in the direction $d = 1$. However, if $T = [1,2]$ then all three functions are uniformly decreasing at $x^* = 0$ in the same direction $d = 1$.

As suggested by the above example, a convex function f is uniformly decreasing at x^* in the direction $d \neq 0$, whenever $\nabla f(x^*, \cdot)$ is continuous and the set

$$E(x^*,d) \triangleq \{t \in T^*: d' \nabla f(x^*,t) = 0\}$$

is empty. Its complement

$$S(x^*,d) = T^* \backslash E(x^*,d) = T^*$$

is then compact. In particular, all analytic functions not identically zero are uniformly decreasing. However, a characterization of optimality for problem (P) with such constraint functions is already given by Corollary 4.

An important uniformity property of convex functions with compact $S(x^*,d)$ follows:

LEMMA 2: Let $f(x,t)$ be convex and differentiable in x, for every t in a compact set $T \subset R^m$, and continuous in t, for every $x \in R^n$. Suppose further that for some x^* and $d \neq 0$ in R^n, the set $S(x^*,d)$ is nonempty and compact. Then there exists $\bar{\alpha} > 0$ such that

(22) $$f(x^* + \alpha d, t) < 0, \quad 0 < \alpha \leqslant \bar{\alpha}$$

for all $t \in S(x^*, d)$.

PROOF: Suppose that such $\bar{\alpha} > 0$ does not exist. Then there exists a sequence $\{t_i\} \subset S(x^*, d)$ and a sequence $\{\alpha_i\}$, $\alpha_i = \alpha_i(t_i) > 0$ such that

$$f(x^* + \alpha_i d, t_i) = 0,$$

$$f(x^* + \alpha d, t_i) < 0, \quad 0 < \alpha < \alpha_i$$

and

(23) $$f(x^* + \alpha d, t_i) > 0, \ \alpha > \alpha_i$$

with $\inf \{\alpha_i\} = 0$. Since $S(x^*, d)$ is compact, $\{t_i\}$ contains a convergent subsequence $\{t_{i_j}\}$. Let $\hat{t} \in S(x^*, d)$ be the limit point of $\{t_{i_j}\}$. Now

$$d' \nabla f(x^*, \hat{t}) < 0$$

implies that there exists $\hat{\alpha} > 0$ such that

$$f(x^* + \alpha d, \hat{t}) < 0, \quad 0 < \alpha \leqslant \hat{\alpha}.$$

In particular,

(24) $$f(x^* + \hat{\alpha} d, \hat{t}) < 0.$$

For any $\epsilon > 0$ there exists $j_o = j_o(\epsilon)$ such that

(25) $$|t_{i_j} - \hat{t}| < \epsilon \text{ and } \alpha_{i_j} < \hat{\alpha} \text{ for all } j > j_o.$$

Now (23) and (25) imply

(26) $$f(x^* + \hat{\alpha} d, t_{i_j}) > 0 \text{ for all } j > j_o.$$

But the inequalities (24) and (26) contradict the continuity of $f(x^* + \hat{\alpha} d, \cdot)$.

□

EXAMPLE 5: Consider again

$$f^2(x, t) = x^2 - tx, \ t \in T = [1, 2].$$

This function is uniformly decreasing at $x^* = 0$ in the direction $d = 1$. The inequality (22) holds for every $0 < \bar{\alpha} < 1$, in particular $\bar{\alpha} = \frac{1}{2}$. If the above interval T is replaced by $\bar{T} = [0, 1]$, then f^2 is not uniformly decreasing at $x^* = 0$ with $d = 1$. An $\bar{\alpha} > 0$ satisfying (22) here does not exist.

A characterization of optimality for programs (P), with constraint functions which have the uniform mean value property and are uniformly decreasing, follows.

THEOREM 3: Let x^* be a feasible solution of problem (P), where f^k, $k \in P^*$ have the uniform mean value property. Suppose also that f^k, $k \in P^*$ are uniformly decreasing at x^* in every feasible direction d. Then x^* is an optimal solution of (P) if, and only if, for every $\alpha^* > 0$ the system

(A) $$d' \nabla f^o(x^*) < 0,$$

(B_4) $$\begin{cases} d' \nabla f^k(x^*, t) < 0 \text{ or } d \in D_k(x^*, t) \\ \qquad\qquad \text{for all } t \in T_k^* \end{cases},$$

$$\text{(C)} \qquad \begin{cases} \dfrac{d' \nabla f^k(x^* + \alpha^* d, t)}{f^k(x^*, t)} \geqslant -\dfrac{1}{\alpha^*} \\ \text{for all } t \in T_k \backslash T_k^*, \\ k \in P^* \end{cases}$$

is inconsistent.

PROOF: Parts (A) and (C) are proved as in the case of Theorem 1. It is left to show that the existence of $\bar{\alpha} > 0$ satisfying (2-a) is equivalent to the consistency of (B_4). It is clear that (2-a) implies (B_4). In order to show that (B_4) implies (2-a) we use the assumption that the functions $\{f^k(x,t)\colon k \in P^*\}$ are uniformly decreasing at x^* in the direction d. When (B_4) holds, then for every $k \in P^*$ there exist $\alpha_k > 0$ and $\alpha_k^o > 0$ such that

$$\text{(27)} \qquad \begin{cases} f^k(x^* + \alpha d, t) < 0, \quad 0 < \alpha \leqslant \alpha_k \\ \text{for all } t \in S_k \triangleq \{t \in T_k^*\colon d' \nabla f^k(x^*, t) < 0\}, \end{cases}$$

by Lemma 2, and

$$\text{(28)} \qquad \begin{cases} f^k(x^* + \alpha d, t) = 0, \quad 0 < \alpha \leqslant \alpha_k^o \\ \text{for all } t \in T_k^* \backslash S_k, \end{cases}$$

since $d \neq 0$. The latter follows by part (ii) of Definition 2 and the convexity of f^k. Let

$$\text{(29)} \qquad \bar{\alpha} \triangleq \min_{k \in P^*} \{\alpha_k, \alpha_k^o\} > 0.$$

Clearly (27) and (28) can be written as the single statement (2-a) with $\bar{\alpha}$ chosen as in (29).

□

The following example shows that the assumption that $\{f^k(x,t)\colon k \in P^*\}$ be uniformly decreasing at x^* cannot be omitted in Theorem 3.

EXAMPLE 6: Consider the program

$$\text{Min } f^o(x) = -x$$

s.t.

$$f(x,t) \leqslant 0, \text{ for all } t \in T = [0,1]$$

where

$$f(x,t) = \begin{cases} t(x-t)^2 & \text{if } x > t \\ 0 & \text{if } x \leqslant t. \end{cases}$$

The feasible set consists of the single point $x^* = 0$, which is therefore optimal. One can verify, after some manipulation, that the constraint function f has the uniform mean value property at x^*. (For every $\alpha > 0$ there exists $0 < \hat{\alpha} \leqslant \frac{1}{2}\alpha$ such that (MV) holds.) However, f is not uniformly decreasing at x^*. In order to demonstrate that Theorem 3 here fails, first we note that $T^* = T = [0,1]$, so the condition (C) is redundant. Since $d = 1$ is in the cone of directions of constancy $D(x^*,t)$ for every $t \in [0,1]$, we conclude that the system (A), (B_4) is here consistent, contrary to the statement of the theorem. Therefore the assumption that the constraint functions be uniformly decreasing cannot be omitted in Theorem 3.

5. THE FRITZ JOHN AND KUHN-TUCKER THEORIES FOR SEMI-INFINITE PROGRAMMING

In contrast to the characterizations of optimality stated in the preceding sections we will now recall the Fritz John and Kuhn-Tucker theories for semi-infinite programming. In the sequel we use the following concept from the duality theory of semi-infinite programming, e.g. [3].

DEFINITION 4: Let I be an arbitrary index set, $\{p^i: i \in I\}$ a collection of vectors in R^m and $\{c_i: i \in I\}$ a collection of scalars. The linear inequality system

$$u^t p^i \leqslant c_i, \text{ for all } i \in I$$

is canonically closed if the set of coefficients $\{((p^i)^t, c_i): i \in I\}$ is compact in R^{m+1} and there exists a point u^o such that

$$(u^o)^t p^i < c_i, \text{ for all } i \in I.$$

We will say that problem (P) is canonically closed at x^* if the system

$$(B_5) \qquad d^t \nabla f^k(x^*,t) \leqslant 0 \text{ for all } t \in T_k^*,\ k \in P^*$$

is canonically closed.

REMARK 3: All constraint functions of problem (P) can have the uniform mean value property, or they can be uniformly decreasing, without problem (P) being canonically closed.

Lemma 3 below is a specialized version of Theorem 3 from [3], adjusted to our need. It is related to the following pair of the semi-infinite linear programs:

(I)	(II)
Inf $u^t p^o$	Sup $\sum_{i \in I} c_i \lambda_i$
s.t.	s.t.
$u^t p^i \geqslant c_i$, all $i \in I$	$\sum_{i \in I} p^i \lambda_i = p^o$
$u \in R^m$	$\lambda \in S,\ \lambda \geqslant 0,$

where S is the vector space of all vectors $\{\lambda_i: i \in I\}$ with only finitely many nonzero entries. Denote by V_I and V_{II} the optimal values of (I) and (II), respectively.

LEMMA 3: Assume that the linear inequality system of problem (I) is canonically closed. If the feasible set of problem (I) is nonempty and V_I is finite, then problem (II) is consistent and $V_{II} = V_I$. Moreover, V_{II} is a maximum.

The concept of a canonically closed system is used in the proof of the dual statement of the following theorem.

THEOREM 4: ("The Fritz John Necessity Theorem") Let x^* be an optimal solution of problem (P). Then the system

$$(A) \qquad d^t \nabla f^o(x^*) < 0$$

(B_1) $\quad d^t \nabla f^k(x^*,t) < 0$ for all $t \in T_k^*$, $k \in P^*$

is inconsistent or, dually, the system

$$(FJ) \quad \begin{cases} \lambda^o \nabla f^o(x^*) + \sum_{k \in P^*} \sum_{t \in T_k^*} \lambda_t^k \nabla f^k(x^*,t) = 0 \\ \lambda^o, \{\lambda_t^k : t \in T_k^*,\ k \in P^*\} \text{ nonnegative scalars,} \\ \text{not all zero and of which only finitely many are positive} \end{cases}$$

is consistent.

PROOF: If x^* is optimal, then the inconsistency of the system (A), (B_1) is well-known, e.g. [4, Lemma 1]. In order to prove the dual statement, we note that the inconsistency of (A), (B_1) is equivalent to $\mu^* = 0$ being the optimal value of the semi-infinite linear program

$(\hat{I})$ $\quad$ Min μ

s.t.

$$d^t \nabla f^o(x^*) + \mu \geqslant 0$$

$$d^t \nabla f^k(x^*,t) + \mu \geqslant 0, \text{ all } t \in T_k^*,\ k \in P^*$$

$$\begin{pmatrix} d \\ \mu \end{pmatrix} \in R^{n+1}.$$

The dual of $(\hat{I})$ is

$(\hat{II})$

Max 0

s.t.

$$\lambda^o \nabla f^o(x^*) + \sum_{k \in P^*} \sum_{t \in T_K^*} \lambda_t^k \nabla f^k(x^*,t) = 0$$

$$\sum_{k \in P^*} \sum_{t \in T_k^*} \lambda_t^k = 1,$$

$\lambda_t^k \geqslant 0$, only finitely many are positive.

The feasible set of problem $(\hat{I})$ is clearly nonempty and canonically closed ($d = 0$, $\mu = 1$ satisfy the constraints of $(\hat{I})$ with strict inequalities). Lemma 3 is now readily applicable to the pair $(\hat{I})$, $(\hat{II})$, which proves (FJ).

□

The dual statement in Theorem 4 is the Fritz John optimality condition for semi-infinite programming. For an equivalent formulation the reader is referred to Gehner's paper [4].

Under various "constraint qualifications" such as Slater's condition:

$$\exists \hat{x} \in R^n \ni f^k(\hat{x},t) < 0 \text{ for all } t \in T_k,\ k \in P$$

or the "Constraint Qualification II" of Gehner [4], one can set $\lambda_o = 1$ in Theorem 4. In fact, the same is possible if problem (P) is canonically closed at x^*, i.e. if there exists $\hat{d}$ such that

(30) $\hat{d}' \nabla f^k(x^*,t) < 0$ for all $t \in T_k^*$, $k \in P^*$.

This is easily seen by multiplying the equation in (FJ) by $\hat{d}$ satisfying (30). Note that the canonical closedness assumption is a semi-infinite version of the Arrow-Hurwicz-Uzawa constraint qualification, e.g. [12]. The latter constraint qualification is implied by Slater's condition.

The Fritz John condition (FJ) with $\lambda_o = 1$ is a semi-infinite version of the Kuhn-Tucker condition, e.g. [12]. While the Fritz John condition is necessary but not sufficient, the Kuhn-Tucker condition is sufficient but not necessary for optimality. If a constraint qualification is assumed, then the Kuhn-Tucker condition is both necessary and sufficient for optimality for problem (P). If a constraint qualification is not satisfied then the Fritz John condition fails to establish the optimality and the Kuhn-Tucker condition fails to establish the nonoptimality of a feasible point x^*. In contrast, our results are applicable. This will be demonstrated by two examples. (See also an example, taken from approximation theory, in Section 6.)

EXAMPLE 7: Consider the semi-infinite convex problem

$$\text{Min } f^o = x_1 - x_2$$

s.t.

$$f^1 = x_1^2 + tx_2 - t^2 \leqslant 0 \text{ for all } t \in T_1 = [0,1]$$

$$f^2 = -x_1 - tx_2 - t \leqslant 0 \text{ for all } t \in T_2 = [0,1].$$

The feasible set is

$$F = \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} : -1 \leqslant x_2 \leqslant 0 \right\}$$

and the optimal solution is $x^* = (0,0)'$. For this point

$$T_1^* = T_2^* = \{0\}, \quad P^* = \{1,2\}.$$

The system (B_5) is

$$0 \leqslant 0$$

$$-d_1 \leqslant 0,$$

obviously not canonically closed. The Kuhn-Tucker condition is

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} + \lambda_1 \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\lambda_1 \geqslant 0, \ \lambda_2 \geqslant 0$$

which clearly fails.

One can easily verify that the constraint functions f^1 and f^2 have the uniform mean value property. Also, these functions are uniformly decreasing at $x^* = 0$ in every direction $d \neq 0$. (The sets T_1^* and T_2^* are singletons!) Thus Theorem 3 is applicable. Conditions (A), (B_4) and (C) are here

(A) $$d_1 - d_2 < 0$$

(B_4) $$\begin{cases} 0 < 0 \text{ or } d_1 = 0, \ d_2 \in R \\ -d_1 \leqslant 0 \end{cases}$$

$$\text{(C)} \qquad \begin{cases} \dfrac{\alpha^* d_1^2 + t d_2}{-t^2} \geqslant -\dfrac{1}{\alpha^*} \text{ for all } t \in (0,1] \\ \dfrac{-d_1 - t d_2}{-t} \geqslant -\dfrac{1}{\alpha^*} \text{ for all } t \in (0,1]. \end{cases}$$

This reduces to

$$\text{(31)} \qquad \begin{aligned} & d_1 = 0, \ d_2 > 0 \\ & \frac{d_2}{-t} \geqslant -\frac{1}{\alpha^*} \text{ for all } t \in (0,1] \\ & -d_2 \geqslant -\frac{1}{\alpha^*}. \end{aligned}$$

Since $d_2 > 0$, the inequality (31) cannot hold for any $\alpha^* > 0$. Hence, by Theorem 3, $x^* = (0,0)'$ is optimal. The optimality of a feasible point is thus established here using Theorem 3 and not by the Kuhn-Tucker condition which here fails.

Consider now the point $x^* = (0,-1)'$. Here

$$T_1^* = \{0\}, \ T_2^* = [0,1], \ P^* = \{1,2\}.$$

It is easy to verify that the Fritz John condition is satisfied in spite of the fact that x^* is not optimal. Conditions (A), (B) and (C) are here

$$\text{(A)} \qquad d_1 - d_2 < 0$$

$$\text{(B)} \qquad \begin{cases} 0 \leqslant 0 \\ -d_1 - t d_2 \leqslant 0 \text{ for all } t \in [0,1] \end{cases}$$

$$\text{(C)} \qquad \frac{\alpha^* d_1^2 + t d_2}{-t - t^2} \geqslant -\frac{1}{\alpha^*} \text{ for all } t \in (0,1].$$

For $\alpha^* = 1$, these conditions are satisfied by $d_1 = 0$, $d_2 = 1$. Hence, by Theorem 1, the point $x^* = (0,1)'$ is not optimal. Both the Fritz John and the Kuhn-Tucker theories fail to characterize optimality in this example because a constraint qualification (or a regularization condition, e.g. [1]) is not here satisfied.

Although the Fritz John and Kuhn-Tucker theories fail to characterize optimality, they can be used to formulate, respectively, either the necessary or the sufficient conditions of optimality.

In the remainder of the section we will show that the ordinary Kuhn-Tucker condition (i.e. the (FJ) condition with $\lambda_o = 1$) can be weakened by assuming an asymptotic form. For a related discussion in Banach spaces the reader is referred to [16].

THEOREM 5: ("The Kuhn-Tucker Sufficiency Theorem") Let x^* be a feasible solution of problem (P). Then x^* is optimal if the system

$$\text{(A)} \qquad d' \nabla f^o(x^*) < 0$$

$$\text{(B}_5\text{)} \qquad d' \nabla f^k(x^*, t) \leqslant 0 \text{ for all } t \in T_k^*, \quad k \in P^*$$

is inconsistent or, dually, if the system

$$(\overline{K-T}) \quad \begin{cases} \nabla f^o(x^*) + \overline{\sum_{k \in P^*} \sum_{t \in T_k^*} \lambda_t^k \nabla f^k(x^*,t)} = 0 \\ \{\lambda_t^k : t \in T_k^*,\ k \in P^*\} \text{ nonnegative scalars} \\ \text{of which only finitely many are positive} \end{cases}$$

is consistent.

PROOF: If the system (A), (B_5) is inconsistent, so is (A), (B). (Recall that $D_k(x^*,t) \subset \{d: d' \nabla f^k(x^*,t) = 0\}$.) Hence, in particular, the system (A), (B), (C) is inconsistent. Following the proof of Theorem 1, one concludes that x^* is optimal. The inconsistency of (A), (B_5) is equivalent to the consistency of $(\overline{K-T})$, by e.g. [11, Corollary 5].

□

REMARK 4: The "asymptotic" form of the Kuhn-Tucker conditions $(\overline{K-T})$ gives a weaker sufficient condition for optimality than the familiar (i.e., without the closure) condition

$$(K-T) \quad \begin{cases} \nabla f^o(x^*) + \sum_{k \in P^*} \sum_{t \in T_k^*} \lambda_t^k \nabla f^k(x^*,t) = 0 \\ \{\lambda_t^k : t \in T_k^*,\ k \in P^*\} \text{ nonnegative scalars} \\ \text{of which only finitely many are positive.} \end{cases}$$

In some situations the primal Kuhn-Tucker conditions (A), (B_5) may be easier to apply than $(\overline{K-T})$. This will be illustrated on the following problem taken from [8, Example 2.4].

EXAMPLE 8: Consider

$$\text{Min } f^o = 4x_1 + \frac{2}{3}(x_4 + x_6)$$

s.t.

$$f^1 = -x_1 - t_1 x_2 - t_2 x_3 - t_1^2 x_4 - t_1 t_2 x_5 - t_2^2 x_6 + 3 - (t_1 - t_2)^2 (t_1 + t_2)^2 \leqslant 0$$

$$\text{for all } t \in T_1 = \left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} : -1 \leqslant t_i \leqslant 1,\ i = 1,2 \right\}.$$

We will show, using the Kuhn-Tucker theory, that $x^* = (3,0,0,0,0,0)'$ is an optimal solution. The optimality of x^* has been established in [8] by a different approach.

First note that here

$$T_1^* = \left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} : t_1 - t_2 = 0 \text{ or } t_1 + t_2 = 0 \right\} \cap T_1.$$

The system (A), (B_5) becomes

$$(A) \qquad 4d_1 + \frac{2}{3} d_4 + \frac{2}{3} d_6 < 0$$

$$(B_5) \qquad -d_1 - t_1 d_2 - t_2 d_3 - t_1^2 d_4 - t_1 t_2 d_5 - t_2^2 d_6 \leqslant 0$$

$$\text{for all } t \in T_1^*.$$

Substitute in (B_5) the following five points of T_1^*:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

This gives

$$\begin{aligned} -d_1 &\leqslant 0 \\ -d_1 - d_2 - d_3 - d_4 - d_5 - d_6 &\leqslant 0 \\ -d_1 - d_2 + d_3 - d_4 + d_5 - d_6 &\leqslant 0 \\ -d_1 + d_2 - d_3 - d_4 + d_5 - d_6 &\leqslant 0 \\ -d_1 + d_2 + d_3 - d_4 - d_5 - d_6 &\leqslant 0. \end{aligned}$$

Multiply the first inequality by ten thirds and each of the remaining four inequalities by one sixth then add all five inequalities. We get

$$-4d_1 \qquad -\frac{2}{3}d_4 \qquad -\frac{2}{3}d_6 \leqslant 0$$

which contradicts (A). Thus the system (A), (B_5) is inconsistent and $x^* = (3,0,0,0,0,0)'$ is optimal, by Theorem 5.

Theorems 1 and 3 suggest that the presently used constraint qualifications for semi-infinite programming problems are too restrictive because they do not employ the topological properties of problem (P), such as the uniform mean value property or the uniformly decreasing constraints.

6. AN APPLICATION TO CHEBYSHEV APPROXIMATION

It is well-known that there is a close connection between convex programming and approximation theory, e.g. [5],[13]. In fact, many approximation problems can be formulated as convex semi-infinite programming problems in which case the results of this paper are readily applicable. In particular, the problem of linear Chebyshev approximation subject to side constraints

(MM)

$$\text{Min} \left(\max_{t \in T} |f(t) - \sum_{i=1}^{n} x_i g_i(t)| \right)$$

s.t.

$$l(t) \leqslant \sum_{i=1}^{n} x_i g_i(t) \leqslant u(t) \text{ for all } t \in T$$

is equivalent to the linear semi-infinite programming problem

(L)

$$\text{Min } x_{n+1}$$

s.t.

$$\left.\begin{aligned} -x_{n+1} \leqslant \sum_{i=1}^{n} x_i g_i(t) - f(t) &\leqslant x_{n+1} \\ l(t) \leqslant \sum_{i=1}^{n} x_i g_i(t) &\leqslant u(t) \end{aligned}\right\} \text{ for all } t \in T.$$

Corollary 3 of this paper can be applied to (L) and it gives a characterization of the best approximation for the problem (MM). Uniqueness of the best approximation can be checked using Theorem 2. Rather than going into details we will illustrate this application by an example.

EXAMPLE 9: The approximation problem stated in this example is taken from [4], see also [15]. It shows that there exist situations when the Kuhn-Tucker theory for semi-infinite programming fails to establish optimum even in the case of linear constraints. However, the optimality is established using the results of this paper.

The linear Chebyshev approximation problem is

$$\text{Min} \left(\max_{t \in [0,1]} |t^4 - x_1 - x_2 t| \right)$$

s.t.

$$-t \leqslant x_1 + x_2 t \leqslant t^2, \text{ for all } t \in [0,1].$$

An equivalent linear semi-infinite programming problem is

$$\text{Min } f^o = x_3$$

s.t.

$$\left.\begin{aligned} f^1 &= t^4 - x_1 - x_2 t - x_3 \leqslant 0 \\ f^2 &= -t^4 + x_1 + x_2 t - x_3 \leqslant 0 \\ f^3 &= -t^2 + x_1 + x_2 t \leqslant 0 \\ f^4 &= -t - x_1 - x_2 t \leqslant 0 \end{aligned}\right\} \text{ for all } t \in [0,1].$$

Is $x^* = (0,0,1)'$ optimal?

Here $T_1^* = \{1\}$, $T_2^* = \emptyset$, $T_3^* = \{0\}$, $T_4^* = \{0\}$ and $P^* = \{1,3,4\}$. The system (A), (B_5) is

(A) $$d_3 < 0$$

(B_5) $$\begin{cases} -d_1 - d_2 - d_3 \leqslant 0 \\ d_1 \leqslant 0 \\ -d_1 \leqslant 0 \end{cases}$$

and it is clearly consistent (set e.g. $d_1 = 0$, $d_2 = 1$, $d_3 = -1$). Therefore, Theorem 5 cannot be applied. (Since the system $(K-T)$ is inconsistent, $x^* = (0,0,1)'$ is not a "Kuhn-Tucker point".) But the system

(A) $$d_3 < 0$$

(B_2) $$\begin{cases} -d_1 - d_2 - d_3 \leqslant 0 \\ d_1 \leqslant 0 \\ -d_1 \leqslant 0 \end{cases}$$

(C_1) $$\begin{cases} \dfrac{-d_1 - d_2 t - d_3}{t^4 - 1} \geqslant -1, \text{ for all } t \in [0,1) \\ \dfrac{d_1 + d_2 t}{-t^2} \geqslant -1, \text{ for all } t \in (0,1] \\ \dfrac{d_1 + d_2 t}{t} \geqslant -1, \text{ for all } t \in (0,1] \end{cases}$$

is inconsistent. (First, $d_1 = 0$, by the last two inequalities in (B_2). Now (A) and (B_2) imply $d_2 > 0$. This contradicts $d_2 \leqslant 0$ obtained from the second inequality in (C_1).) Therefore $x^* = (0, 0, 1)'$ is optimal, by Corollary 1.

ACKNOWLEDGMENT

The authors are indebted to Professor G. Schmidt for providing some of the constraint functions used in Examples 1, 3 and 4, Mr. H. Wolkowicz for providing a counter-example to one of their earlier conjectures and the referee for his recommendations about organization of the paper and providing a correct version of Lemma 3.

REFERENCES

[1] Ben-Tal. A., A. Ben-Israel and S. Zlobec, "Characterization of Optimality in Convex Programming without a Constraint Qualification," Journal of Optimization Theory and Applications *20*, 417-437 (1976).

[2] Charnes, A., W.W. Cooper and K.O. Kortanek, "Duality, Haar Programs and Finite Sequence Spaces," Proceedings of the National Academy of Science, *48*, 783-786 (1962).

[3] Charnes, A., W.W. Cooper and K.O. Kortanek, "On the Theory of Semi-Infinite Programming and a Generalization of the Kuhn-Tucker Saddle Point Theorem for Arbitrary Convex Functions," Naval Research Logistics Quarterly, *16*, 41-51 (1969).

[4] Gehner, K.R., "Necessary and Sufficient Conditions for the Fritz John Problem with Linear Equality Constraints," SIAM Journal on Control, *12*, 140-149 (1974).

[5] Gehner, K.R., "Characterization Theorems for Constrained Approximation Problems via Optimization Theory," Journal of Approximation Theory, *14*, 51-76 (1975).

[6] Gorr, W. and K.O. Kortanek, "Numerical Aspects of Pollution Abatement Problems: Constrained Generalized Moment Techniques," Carnegie-Mellon University, School of Urban and Public Affairs, Institute of Physical Planning Research Report No. 12 (1970).

[7] Gustafson, S.A. and K.O. Kortanek, "Analytical Properties of Some Multiple-Source Urban Diffusion Models," Environment and Planning, *4*, 31-41 (1972).

[8] Gustafson, S.A. and K.O. Kortanek, "Numerical Treatment of a Class of Semi-Infinite Programming Problems," Naval Research Logistics Quarterly, *20*, 477-504 (1973).

[9] Gustafson, S.A. and J. Martna, "Numerical Treatment of Size Frequency Distributions with Computer Machine," Geologiska Foreningens Forhandlingar, *84*, 372-389 (1962).

[10] Kantorovich, L.V. and G.Sh. Rubinshtein, "Concerning a Functional Space and Some Extremum Problems," Dokl. Akad. Nauk. SSSR *115*, 1058-1061 (1957).

[11] Lehmann, R. and W. Oettli, "The Theorem of the Alternative, the Key-Theorem and the Vector-Maximum Problem," Mathematical Programming, *8*, 332-344 (1975).

[12] Mangasarian, O., *Nonlinear Programming*, McGraw Hill, New York (1969).

[13] Rabinowitz, P., "Mathematical Programming and Approximation," *Approximation Theory*, A. Talbot (editor), Academic Press (1970).

[14] Rockafellar, R.T., *Convex Analysis*, Princeton University Press, Princeton, N.J. (1970).

[15] Taylor, G.D., "On Approximation by Polynomials Having Restricted Ranges," Journal on Numerical Analysis, *5*, 258-268 (1968).

[16] Zlobec, S., "Extensions of Asymptotic Kuhn-Tucker Conditions in Mathematical Programming," SIAM Journal on Applied Mathematics, *21*, 448-460 (1971).

SOLVING INCREMENTAL QUANTITY DISCOUNTED TRANSPORTATION PROBLEMS BY VERTEX RANKING

Patrick G. McKeown

University of Georgia
Athens, Georgia

ABSTRACT

Logistics managers often encounter incremental quantity discounts when choosing the best transportation mode to use. This could occur when there is a choice of road, rail, or water modes to move freight from a set of supply points to various destinations. The selection of mode depends upon the amount to be moved and the costs, both continuous and fixed, associated with each mode. This can be modeled as a transportation problem with a piecewise-linear objective function. In this paper, we present a vertex ranking algorithm to solve the incremental quantity discounted transportation problem. Computational results for various test problems are presented and discussed.

1. INTRODUCTION

Whenever a logistics manager is making a decision about the movement of freight, he is often faced with choosing from among different modes of transportation. Movement of freight by air or motor express may involve no fixed costs to the transporter, but will usually involve relatively higher variable costs than either rail or water. However, both rail and water can involve the investment of large sums for rail sidings or docking facilities. The problem of selecting freight modes can be modeled as a transportation problem with a piecewise-linear objective function. This problem has been termed the incremental quantity discounted transportation problem, since it is assumed that the variable costs decrease as the amount shipped increases. This comes about due to the lower variable costs for rail or water modes relative to air or road freight costs. The presence of fixed costs for the use of rail or water determines the range of shipment levels over which each cost will be applicable. Figure 1 shows this type of objective function.

In this paper we will present a vertex ranking algorithm to solve this type problem along with the computational results from various sizes and types of problems. Background material is discussed in Section 2, while the details of the algorithm are given in Section 3. An example is worked out in Section 4 while Section 5 gives computational results.

2. BACKGROUND MATERIAL

The incremental quantity discounted transportation problem is a member of a general class of math programming problems, i.e., the piecewise-linear programming problem. Vogt and Even [15] considered the case of the piecewise-linear transportation problem derived from U.S. freight rates. This problem is neither convex nor concave, and has sections of the objective function which are flat or "free." Figure 2 shows this case. Vogt and Evans used separable

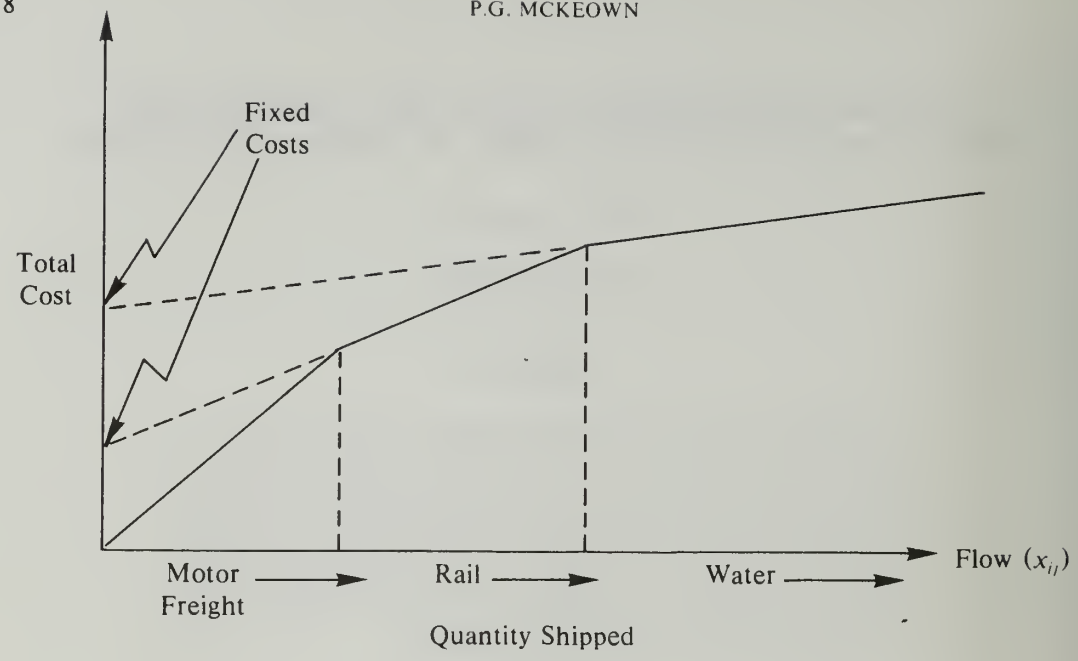


FIGURE 1

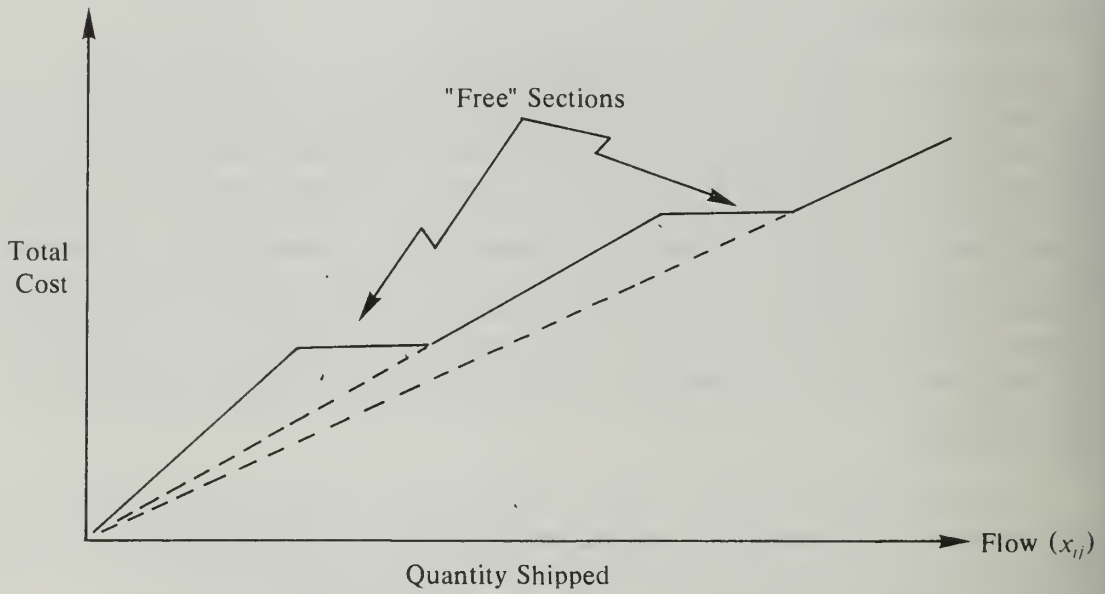


FIGURE 2

nonconvex programming to reach an approximately optimal solution to this problem. Balachandran and Perry [1] consider another version of this problem which they termed the all unit quantity discount transportation problem. The main difference between this and the previous case is the lack of the flat section of the objective function. The latter case is typical of some foreign freight rates, and is shown in Figure 3 below.

Problems similar to this one have been mentioned in the plant location literature, e.g., Townsend [14], and Efroymson and Ray [5]. In these cases, it is suggested that the problem be solved by considering multiple plants, one for each range of demand.

Balachandran and Perry presented a branch and bound algorithm for the all unit quantity discount problem, which they show, will also work for the incremental quantity discount problem as well as fixed charge transportation problems. However, no computational results are

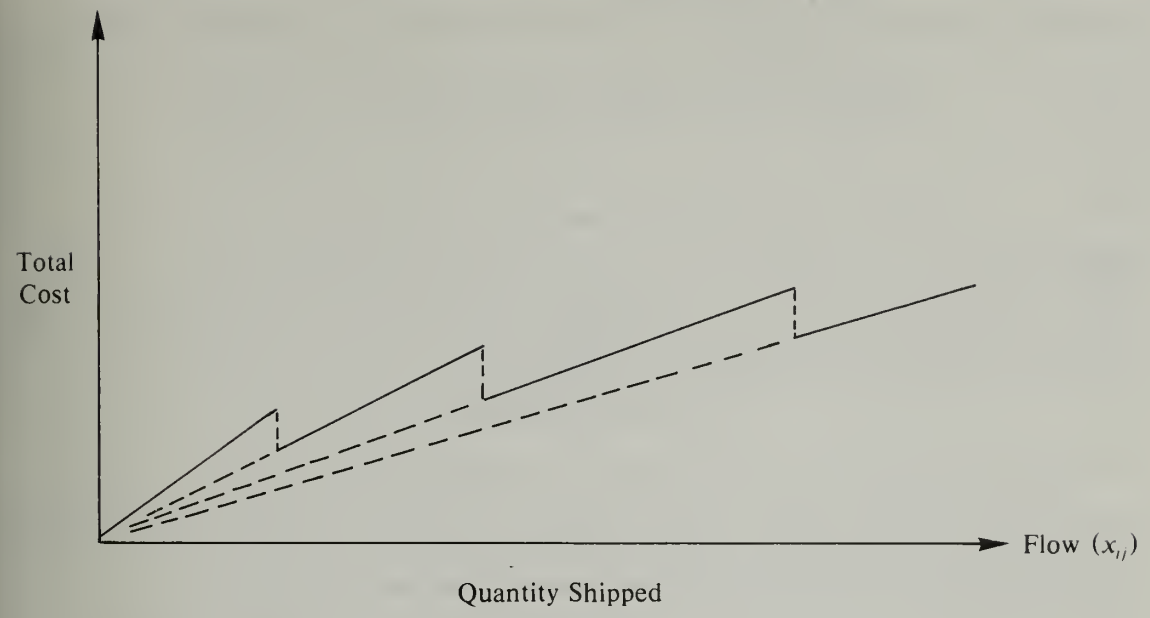


FIGURE 3

given to demonstrate the efficiency of this algorithm. Here, we will consider a vertex ranking algorithm for only the incremental quantity discount problem for two reasons. First, fixed charge transportation problems have been handled in several other places in a manner that has been shown to be superior to vertex ranking [2,8]. Secondly, the incremental quantity discount transportation problem has a concave objective function; but, neither the problem considered by Vogt and Evans, or the all unit quantity discount transportation problem, have nonconcave objective functions. This is crucial to the use of vertex ranking since this procedure will only consider vertices of the constraint set, and the optimal solution to problems with nonconcave objective functions need not occur at a vertex.

The incremental quantity discount problem may be formulated as follows (following the model proposed by Balachandran and Perry [1]):

$$\text{(1)} \qquad \text{Min } \bar{Z} = \sum_i \sum_j C_{ij}^k x_{ij} + \sum_i \sum_j f_{ij}^k y_{ij}^k$$

$$\text{(2)} \qquad \text{subject to } \sum_i x_{ij} = a_i \text{ for } i \in I$$

$$\text{(3)} \qquad \sum_j x_{ij} = b_j \text{ for } j \in J$$

$$\text{(4)} \qquad C_{ij}^k = \begin{cases} C_{ij}^1 & \text{if } \lambda_{ij}^0 \leqslant x_{ij} < \lambda_{ij}^1 \\ C_{ij}^2 & \lambda_{ij}^1 \leqslant x_{ij} < \lambda_{ij}^2 \\ \cdot & \\ \cdot & \\ \cdot & \\ C_{ij}^r & \text{if } \lambda_{ij}^{r-1} \leqslant x_{ij} < \lambda_{ij}^r \leqslant \infty, \end{cases}$$

$$y_{ij} = \begin{cases} 1 \text{ if } \lambda_{ij}^{k-1} \leqslant x_{ij} < \lambda_{ij}^{k} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

$$f_{ij}^{k} = \left[\sum_{v=1}^{k-1} C_{ij}^{v} \left(\lambda_{ij}^{v} - \lambda_{ij}^{v-1} \right) \right] - C_{ij}^{k} \lambda_{ij}^{k-1} \text{ for } k = 2, 3, \ldots, r \tag{6}$$

and

$$f_{ij}^{1} = 0, \; x_{ij} \geqslant 0 \text{ for } \textit{all } i \in 1 \text{ and } j \in J, \tag{7}$$

where

$J = \{1, \ldots, n\}$ = set of sinks,

$I = \{1, \ldots, m\}$ = set of sources,

$R = \{1, \ldots, r\}$ = set of cost intervals.

As may be easily seen, this is a generalization of the fixed charge transportation problem, (see [1]), with a fixed charge, f_{ij}^{k}, and a continuous cost, C_{ij}^{k}, for each range of shipment between source i and destination j. Since the situation which we are attempting to model, i.e., the choice of shipment mode, does involve various levels of fixed charge, (1) - (7) is the proper formulation for this problem. It should be noted that we are implicitly assuming that

$$C_{ij}^{k} \geqslant C_{ij}^{k+1} \text{ for all } i \in I, \; j \in J.$$

This is necessary for the concavity of the objective function. However, we would expect that lower continuous costs would occur for higher shipment levels.

Balachandran and Perry [1] suggested that (1) - (7) may be solved by a branch and bound algorithm. Their procedure is similar to that used to solve travelling salesmen problems by driving out subtours [13]. They solve the transportation problem with all costs set to their lowest value, i.e., C_{ij}^{r}. If any routes have flow below λ_{ij}^{r-1}, branching is done on one of these variables. Two branches are used. Our branch forces the flow over the arc above the lower limit for the cost level C_{ij}^{r}, i.e., $X_{ij} \geqslant \lambda_{ij}^{r-1}$. In the other branch, the infeasible cost, C_{ij}^{r}, is replaced by the feasible cost, C_{ij}^{k}. This continues until a solution is found where the arc flows match the costs used. This is the optimal solution. However, the effectiveness of the procedure is unknown since the authors did not provide any computational results.

It would also appear that the work of Kennington [8] on the fixed charge transportation problem could possibly be modified to solve this problem by having multiple arcs between each set of nodes. Each arc would be bounded by $\lambda_{ij}^{k-1}$and λ_{ij}^{k} with multiple continuous costs and fixed costs. However, this would lead to effectively larger problems, e.g., a problem with 60 arcs and five breakpoints would have 300 variables in the new problem.

3. SOLUTION PROCEDURE

Using the formulation of the incremental quantity discount transportation problem given in (1) - (7), along with the assumption of decreasing costs, we have a problem with linear constraints and concave objective function. It is well known [7] that an optimal solution for problems of this type will occur at a vertex of the constraint set. Examples of other problems that share this condition are the fixed charge problem, the quadratic transportation problem, and the quadratic assignment problem. Murty [12] was the first to suggest a vertex ranking scheme for a problem of this category. He showed that the fixed charge problem could be solved by ranking the vertices of the constraint according to the objective value up to some upper bound. At that point, the optimal solution would be found at one or more of the ranked vertices.

We may formulate any problem with concave objective function and linear constraint as below:

(8) $\quad \text{Min } f(x)$

(9) $\quad s.t.\ x \in S$

(10) $\quad$ where $S = \{x | Ax = b, x \geqslant 0\}$.

Since no "direct" optimization techniques exist for the case where $f(x)$ is nonlinear, we shall look at a procedure for searching the vertices of S. To do this, we will use a linear underapproximation of $f(x)$, say $L(x)$, such that $L(x) \leqslant f(x)$, $x \in S$. In this case, to show that x^* is an optimal solution to (8) - (10), we need only rank the vertices of S until a vertex of x° is found such that $L(x^\circ) \geqslant f(x^*)$. At this point, all vertices that could possibly be optimal have been ranked. This is proved by Cabot and Francis [3].

In order to rank the extreme points of S, we need to use a result also first proved by Murty as Theorem 1 below:

THEOREM 1: If $E_1, E_2, \ldots, E_K$ are the first K vertices of a linear underapproximation problem which are ranked in nondecreasing order according to their objective value, then vertex E_{K+1} must be adjacent to one of $E_1, E_2, \ldots, E_K$.

Simply put, this says that vertex 2 will be adjacent to the optimal solution to the linear underapproximation and vertex 3 will be adjacent to vertex 1, or vertex 2, and so on. This, then, gives us a procedure for ranking the vertices if all adjacent vertices can be found. It is this "if" that can cause problems. These problems arise due to the possibility of degeneracy in S. If S is degenerate, then there may exist multiple bases for the same vertex. This implies that *all* such bases must be available before one can be sure that all adjacent vertices have been found. Finding all such bases for finding and "scanning" all adjacent vertices can be quite cumbersome. However, a recent application of Chernikova's work [4,9] has been shown to be a way around the problem of degeneracy.

Vertex ranking has been used by McKeown [10] to solve fixed charge problems and Fluharty [6] to quadratic assignment problems. Cabot and Francis [3] also proposed the use of vertex ranking to solve a certain class of nonconvex quadratic programming problems, e.g., quadratic transportation problems. For a survey of vertex ranking procedures, see [11].

In our problem, we need to determine the linear underapproximation to the first objective function, (1). We may do this by first noting that

(11) $\quad \mu_{ij} = \min \{a_i, b_j\}$

is an upper bound on x_{ij}. We may then note that if $F(x_{ij}) = C_{ij}^k x_{ij} + f_{ij}^k y_{ij}^k$ then

$$l_{ij} = \frac{F(u_{ij}) - F(0)}{u_{ij}}$$

(12) $$= \frac{C_{ij}^k u_{ij} + f_{ij}^k}{u_{ij}}$$

for $\lambda_{ij}^{k-1} \leqslant u_{ij} \leqslant \lambda_{ij}^k$, is a linear underapproximation to $F(x_{ij})$.

We may now form a problem to rank vertices, i.e.,

(13) $$\text{Min } Z = \sum_i \sum_j \left(\frac{C_{ij}^k u_{ij} + f_{ij}^k}{u_{ij}} \right) x_{ij} = \sum_i \sum_j l_{ij} x_{ij}$$

subject to (2) - (7)

for $\lambda_{ij}^{k-1} \leqslant u_{ij} < \lambda_{ij}^k$.

Using (13) and (2) - (7), we may rank vertices as discussed earlier until some vertex x° is found such that $L(x^\circ) = \sum_i \sum_j l_{ij}x_{ij} \geqslant f(x^*)$ where x^* is a candidate for optimality. We may start with x^* equal to the optimal solution to (13) and (2) - (7), and then update it as new, possibly better solutions to (1) - (7), are found by the ranking procedure. When all vertices x are found such that $L(x) < f(x^*)$, the solution procedure terminates with the present candidate being optimal.

EXAMPLE: As an example of our procedure, we will solve an incremental quantity discount version of the example problem presented by Balachandran and Perry [1]. Table 1 below gives the supplies, demands, and costs, for each range of shipment. Table 2 gives the optimal solution to the linear underapproximation problem. The values of l_{ij} are given in the upper right-hand corner of each cell with shipment being circled in the basic cells.

TABLE 1

Source \ Destination	1	2	3	4	Warehouse Capacity
1	$3[20 \leqslant x_{11} < \infty]$ $4[10 \leqslant x_{11} < 20]$ $5[0 \leqslant x_{11} < 10]$	$6[10 \leqslant x_{12} \leqslant \infty]$ $7[5 \leqslant x_{12} < 10]$ $8[0 \leqslant x_{12} < 5]$	$3[27 \leqslant x_{13} < \infty]$ $4[15 \leqslant x_{13} < 27]$ $5[5 \leqslant x_{13} < 15]$	One price bracket 4	80
2	One price bracket 6	$5[65 \leqslant x_{22} < \infty]$ $6[20 \leqslant x_{22} < 65]$ $8[0 \leqslant x_{22} < 20]$	$8[10 \leqslant x_{23} \leqslant \infty]$ $9[5 \leqslant x_{23} < 10]$ $10[0 \leqslant x_{23} < 5]$	One price bracket 15	90
3	$1[27 \leqslant x_{31} \leqslant \infty]$ $2[20 \leqslant x_{31} < 27]$ $3[0 \leqslant x_{31} < 20]$	$3[60 \leqslant x_{32} < \infty]$ $4[30 \leqslant x_{32} < 60]$ $5[0 \leqslant x_{32} < 30]$	$10[20 \leqslant x_{33} \leqslant \infty]$ $11[10 \leqslant x_{33} < 20]$ $12[0 \leqslant x_{33} < 10]$	$5[30 \leqslant x_{34} \leqslant \infty]$ $6[20 \leqslant x_{34} < 30]$ $7[0 \leqslant x_{34} < 20]$	55
Market Demand S	70	60	35	60	

TABLE 2

Source \ Destination	1	2	3	4	Warehouse Capacity
1	3.43	6.25	4.20 (20)	4.00 (60)	80
2	6.00 (15)	6.67 (60)	8.43 (15)	15.00	90
3	1.85 (55)	4.55	10.86	5.91	55
Market Demands	70	60	35	60	

As an example of the calculation of the l_{ij} values, we will look at l_{11}. First, it is necessary to calculate f_{11}^2 and f_{11}^3 using (6). We will do f_{11}^2.

$$f_{11}^2 = C_{11}^1(\lambda_{11} - \lambda_{11}^0) - C_{11}^2 \cdot \lambda_{11}^1$$
$$= (5)(10) - 4(10) = 10.$$

Similarly, $f_{11}^3 = 30$. $u_{11} = \min\{80, 70\} = 70$. Then, $l_{11} = \dfrac{(3)(70) + 30}{70} = 3.43$.

Now, if we solve this continuous transportation problem, we get a value of $Z = 1042.20$ with the circled cells being basic. If we compute the feasible value for this solution, $\bar{Z} = 1067$. Call this solution X^1.

Now, since this solution is nondegenerate, we may use simplex pivoting to look at each nonbasic cell. The values of these adjacent vertices are given below:

Vertex	Z-Value
(1,1)	1067.10
(1,2)	1118.40
(2,4)	1143.75
(3,2)	1154.40
(3,3)	1141.05
(3,4)	1069.80

Since the Z-value for each vertex is greater than the present value of $\bar{Z}$, we do not need to rank any other vertices, and $\bar{Z}$=1067.0 is the optimal solution value.

4. COMPUTATIONAL RESULTS

To test the vertex ranking procedure discussed here, randomly-generated problems were run. These problems were generated by first generating supplies and demands uniformly between upper and lower bounds, U and L. These supplies and demands were generated so that they were all multiples of 5. This was done to insure the presence of degeneracy in some of the problems. All problems were set up to have discount ranges at 20, 50, 300, 1000, and 2000. By proper selection of L and U, various numbers of ranges could be tested.

The costs for each arc were generated by randomly generating mileages between each set of nodes, and then, inputting discount cost-per-mile values for each range of flow, e.g., 10, 9, 8, etc. The final discount costs were found by multiplying the mileage between each arc times the various costs. In this way, various supply-demand discount ranges and cost configurations could be tested. These problems were generated and solved using a computer code in FORTRAN run on the CYBER 70/74 using the FTN Compiler with OPT = 1.

The problem characteristics and test results are given in Table 1. The second column gives the solution time in seconds, while the third column shows the number of vertices of the linear underapproximation other than the optimal solution that were ranked to solve each problem. The fourth column gives the size of the problem ($m \times n$); the fifth column gives the number of cost ranges that the arc flows would cover; the sixth column gives the cost per mile for each range of flow, p_{ij}^k; the seventh column gives the lower and upper ranges used to generate the supplies and demands; and finally, the last column gives the ranges used to generate mileages. The C_{ij}^k values were determined by $C_{ij}^k = p_{ij}^k$ (mileage). As can be seen, the algorithm successfully solved all problems tested. The most difficult problems were those with three ranges and supplies/demands between 5 and 100. Problems 6 and 13 are identical, except that 6 is over only 3 ranges, while 13 is over 5; but, problem 13 is solved in much less time. In fact, the linear underapproximating transportation problem was found to be optimal and no other extreme points were even ranked. This was also the case in problems 7, 9, 10, 11, and

12, even though the number of variables increased markedly. It is also interesting to note the effect of costs in problems 5, 6, and 7. These are essentially the same problem, but with the present decrease in cost for increasing flow being less in each case. The results are as expected since in problem 7 the linear underapproximation will be closer to the actual objective function than in problems 4 and 5.

TABLE 3 — *Computational Results*

Problem Number	Vertices Ranked	Solution Time	$m \times n$	Number of Ranges	p_{ij}^k	U,L	Mileage Range
1	13	2.604	6 × 8	3	10,9,8	1,50	100,200
2	39	10.289	8 × 8	3	10,9,8	1,50	100,200
3	39	34.103	9 × 9	3	10,9,8	1,50	100,200
4	257	42.938	6 × 8	3	10,9,8	1,100	100,200
5	247	39.799	6 × 8	3	20,18,17	1,100	100,200
6	84	13.353	6 × 8	3	20,19,18	1,100	100,200
7	0	.121	4 × 6	5	20,19,18,17,16	400,500	50,100
8	18	.888	4 × 8	5	20,19,18,17,16	400,500	50,100
9	0	.196	6 × 8	5	20,19,18,17,16	400,500	50,100
10	0	.393	8 × 8	5	20,19,18,17,16	400,500	50,100
11	0	.518	9 × 9	5	20,19,18,17,16	400,500	50,100
12	0	.130	4 × 6	5	10,9,8,7,6	400,500	50,100
13	0	.213	6 × 8	5	20,19,18,17,16	400,500	100,200

It would appear from these results that vertex ranking does hold promise as a solution procedure for incremental cost discount transportation problems. Neither size of problem nor degeneracy appears to have any effect on solution time but cost patterns and number of cost ranges do seem to have a marked effect.

Extensions of this work could be used to solve other concave linear programming problems. Walker [16] discusses the fact that these can be considered as generalizations of fixed charge problems. The main difference would be that the first linear portion would have a positive fixed charge rather than zero, as in the problem discussed here. However, this would not change the approach to the solution used here.

REFERENCES

[1] Balachandran, V. and A. Perry, "Transportation Type Problems with Quantity Discounts," Naval Research Logistics Quarterly, *23*, 195-209 (1976).

[2] Barr, R.L., "The Fixed Charge Transportation Problem," presented at the Joint National Meeting of ORSA/TIMS in Puerto Rico (Nov. 1974).

[3] Cabot, A.V. and R.L. Francis, "Solving Certain Nonconvex Quadratic Minimization Problems by Ranking the Extreme Points," Operations Research *18*, 82-86 (1970).

[4] Chernikiva, N.V., "Algorithm for Finding a General Formula for the Non-negative Solutions of a System of Linear Inequalities," U.S.S.R. Computational Mathematics and Mathematical Physics.

[5] Efroymson, M.A. and T.L. Ray, "A Branch-Bound Algorithm for Plant Location," Operations Research, *14*, 361-368 (1966).

[6] Fluharty, R., "Solving Quadratic Assignment Problems by Ranking the Assignments," unpublished Master's Thesis, Ohio State University (1970).

[7] Hirsch, W.M. and A.J. Hoffman, "Extreme Varieties, Concave Functions, and The Fixed Charge Problem," Communications on Pure and Applied Mathematics, *14*, 355-370 (1961).

[8] Kennington, J.L., "The Fixed Charge Transportation Problem: A Computational Study with a Branch and Bound Code," AIIE Transactions, *8* (1976).

[9] McKeown, P.G. and D.S. Rubin, "Adjacent Vertices on Transportation Polytopes," Naval Research Logistics Quarterly, *22*, 365-374 (1975).

[10] McKeown, P.G., "A Vertex Ranking Procedure for Solving the Linear Fixed Charge Problem," Operations Research *23*, 1183-1191 (1975).

[11] McKeown, P.G., "Extreme Point Ranking Algorithms: A Computational Survey," Proceedings of Bicentennial Conference on Mathematical Programming (1976).

[12] Murty, K., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Operations Research, *16*, 268-279 (1968).

[13] Shapiro, D., "Algorithms for the Solution of the Optimal Cost Travelling Salesmen Problem," Sc.D. Thesis, Washington University, St. Louis (1966).

[14] Townsend, W., "A Production Stocking Problem Analogous to Plant Location," Operations Research Quarterly, *26*, 389-396 (1975).

[15] Vogt, L. and J. Evan, "Piecewise Linear Programming Solutions of Transportation Costs as Obtained from Rate Traffic," AIIE Transactions, *4* (1972).

[16] Walker, Warren E., "A Heuristic Adjacent Extreme Point Algorithm for the Fixed Charge Problem," Management Science, *22*, 587-596 (1976).

AUXILIARY PROCEDURES FOR SOLVING LONG TRANSPORTATION PROBLEMS

J. Intrator and M. Berrebi

Bar-Ilan University
Ramat-Gan, Israel

ABSTRACT

An efficient auxiliary algorithm for solving transportation problems, based on a necessary but not sufficient condition for optimum, is presented.

In this paper a necessary (but not sufficient) condition for a given feasible solution to a transportation problem to be optimal is established, and a special algorithm for finding solutions which satisfy this condition is adapted as an auxiliary procedure for the MODI method.

Experimental results presented show that finding an initial solution which satisfies this necessary condition for problems with $m << n$ eliminates 70%-90% of the MODI iterations. (See Table 1)

TABLE 1 — *Matrix of Principal Results*

m \ n	20	30	40	50	100	200	300
4	0.65	0.69	0.72	0.74	0.88	0.91	0.93
5	0.61	0.67	0.69	0.71	0.84	0.87	0.90
6	0.59	0.65	0.66	0.68	0.80	0.82	0.85
8	0.61	0.62	0.64	0.66	0.76	0.80	0.82
10	0.57	0.65	0.66	0.69	0.73	0.77	0.80
20	0.25	0.27	0.31	0.36	0.45	0.50	0.52

Fraction of Modi iteration eliminated by using the method presented in this paper.

The case when our algorithm is used during the solution process (especially for $m \sim n$) is presently being examined. Our auxiliary procedure requires relatively little computational effort in finding the appropriate candidate for the basis, eliminating entirely the need to calculate the dual variables. It works with positive variables associated with one pair of rows at a time using only the prices of these rows.

Once a loop for any given pair of rows is determined it may be used to insert numerous non-basic cells in these two rows to the basis. The result is a considerable time reduction in determining loops.

The storage and time requirements for the special lists needed in our auxiliary algorithm are fully discussed in [1]. A rigorous proof presented in [1] shows that updating these lists requires no more than $0(m \log n)$ computer operations per MODI iteration.

A Linear Programming Transportation Problem is characterized by a cost matrix C and two positive requirement vectors a and b such that $\sum_i a_i = \sum_j b_j$. The problem is to minimize $\sum_i \sum_j C_{ij}x_{ij}$ subject to

$$\sum_i x_{ij} = b_j \qquad j = 1, 2 \ldots . n$$

$$\text{(A)} \quad \sum_j x_{ij} = a_i \qquad i = 1, 2 \ldots m$$

$$x_{ij} \geqslant 0 \qquad \text{for all } (i,j).$$

A proper perturbation of our problem ensures that:

(1) each feasible basic solution of (A) contains exactly $m + n - 1$ positive variables x_{ij},

(2) corresponding to each nonbasic cell (i,j) $(x_{ij} = 0)$ there exists a unique loop of different cells, say $L(i,j) = (i,j_1)\ (i_2,j_1)\ (i_2,j_2)\ (i_3,j_2) \ldots .$

$$\text{(B)} \quad (i_r,j_{r-1})\ (i_r,j)\ (i,j)$$

which contains at most two cells in each row and column, where the cell (i,j) is the unique nonbasic cell,

(3) there are no loops which contain basic cells only.

Notation: For fixed l,k $(1 \leqslant l \neq k \leqslant m)$ we denote

$$V_l = \{j | x_{lj} > 0\} \qquad 1 \leqslant j \leqslant n$$

$$V_{lk} = V_l \cap V_k = \{j | x_{lj} > 0,\ x_{kj} > 0\}.$$

With no loss of generality it is assumed that for each l, $(1 \leqslant l \leqslant m)$ there exists at least one destination (column) $j \in V_l$ such that (l,j) is the unique basic cell of column j. Otherwise, an artificial destination, say J with $x_{lj} = b_j = \epsilon$ where ϵ is an infinitely small positive number will be introduced.

It is easy to see that the feasible solution of the augmented problem of dimension $mx(n + 1)$ satisfies 1), 2), 3) mentioned above.

DEFINITION 1: A destination with a unique basic cell will be called a fundamental destination.

The unique nonbasic cell (i,j) of $L(i,j)$ will be considered for convenience to be the last cell of $L(i,j)$. For each loop $L(i,j)$, say loop (B), we introduce the notation:

$$\text{(C)} \quad C_{L(i,j)} = C_{i_1j_1} - C_{i_2j_1} + C_{i_2j_2} - C_{i_3j_2} + \ldots + C_{i_rj_r} - C_{ij}.$$

It is well to know that

$C_{L(i,j)} = u_i + v_j - C_{ij}$ where u_i and v_j are the dual prices.

DEFINITION 2: A loop with $C_L > 0$ is called an improving loop.

DEFINITION 3: Let l,k be a fixed pair of numbers so that $1 \leqslant l \neq k \leqslant m$, we define

$$A_{lk} = \{j \mid x_{lj} > 0, \quad x_{kj} = 0\} = V_l - V_{lk}$$

$$D_{lk}(j) = C_{lj} - C_{kj} \qquad j = 1, 2, \ldots, n.$$

THEOREM 1: The number of elements in V_{lk} is at most 1.

PROOF: Suppose that $J_1, J_2 \in V_{lk}$ $(1 \leqslant J_1 \neq J_2 \leqslant n)$ then the loop (l,J_1) (k,J_1) (k,J_2) (l,J_2) is of only basic cells contradicting (3) above.

Let J_1 be a fundamental destination of A_{lk}, the purpose of Theorem 2 and Theorem 3 is to show that all the simplex loop $L(i,J)$ and the numbers $C_{L(i,J)}$ $i = l,k$; $J \in A_{lk} \cup A_{kl}$ are determined after the simplex loop $L(k,J)$ is found.

THEOREM 2: $C_{L(k,J_2)} - D_{lk}(J_2) = C_{L(k,J_1)} - D_{lk}(J_1)$ for all $J_2 \in A_{lk}$
J_1 being the above fundamental destination of A_{lk}.

PROOF:

CASE (a) $V_{lk} \neq \phi$. Denote by J the unique member of V_{lk} (Th.1). We have $j \neq J_1$; $j \neq J_2$ (J_1 and $J_2 \notin V_k$) and

$$L(k,J_1) = (k,j)\ (l,j)\ (l,J_1)\ (k,J_1)$$

$$L(k,J_2) = (k,j)\ (l,j)\ (l,J_2)\ (k,J_2)$$

e.g.

$$C_{L(k,J_2)} - D_{lk}(J_2) = C_{L(k,J_1)} - D_{lk}(J_1).$$

CASE (b) $V_{lk} = \phi$. Let $L(k,J_1)$ be the loop (B). Note that $i_r = l$ (since column J_1 contains a basic cell in row l exclusively) and $r > 2$. Otherwise, $i_1 = i_2 = l$ and $L(k,J_1) = (k,j)$ (l,j) (l,J_1) (k,J_1) means that $j \in V_{lk}$ contradicting the fact that $V_{lk} = \phi$.

Consider the loop:

$$L = (k,j_1)\ (i_2,j_1)\ (i_2,j_2)\ (i_3,j_2)\ \ldots\ldots\ (l,j_{r-1})\ (l,J_2)\ (k,J_2),\ (i_1 = k)$$

obtained from $L(k,J_1)$ by substituting J_2 for J_1. Let us show that either $L(k,J_2) = L$ or $L(k,J_2)$ can be obtained from L by deleting two identical cells.

At first, observe that all rows and columns of L (except perhaps J_2) contain exactly two different cells of L. The column J_2 has not appeared previously (unless $J_2 = j_{r-1}$) because it equals one of the previous members j_s, $1 \leqslant s \leqslant r-2$, then the loop (i_{s+1},j_s) (i_{s+1},j_{s+1}) $\ldots$ (i_r,J_2) will be a loop of basic cells only which contradicts (3).

Thus, only two possibilities exist:

(1) $J_2 \neq j_{r-1}$ and $L(k,J_2) = L$ or

(2) $J_2 = j_{r-1}$ and $L(k,J_2)$ is obtained from L by deleting the two identical cells (l,j_{r-1}) and (l,J_2).

Since this deleting does not effect the value of $C_{L(k,J_2)}$, we have for both possibilities

$$C_{L(k,J_2)} - D_{lk}(J_2) = C_{L(k,J_1)} - D_{lk}(J_1).$$

THEOREM 3: Let J_1 and J_2 be the destinations defined in Theorem 2 and $J_3 \in A_k$. We shall prove that

$$C_{L(l,J_3)} = -\,[C_{L(k,J_2)} - D_{lk}(J_2)] + D_{kl}(J_3).$$

PROOF: Let $L(k,J_1)$ be the loop (B) with $i_r = l$ because J is fundamental.

Consider the loop L defined by

$$L = (i_r,j_{r-1})\ (i_{r-1},\ j_{r-1}) \ldots (i_2,j_1)\ (i_1,j_1)\ (k,J_3)\ (l,J_3).$$

CASE (a) $V_{lk} \neq \phi$. Same proof as in Theorem 2.

CASE (b) $V_{lk} = \phi$. By the same argument as in Theorem 2 we can show that there are only two possibilities.

1) $J_3 \neq j_1$ which implies that $L(l,J_3) = L$.

2) $J_3 = j_1$. In this case $(r > 2)$ and $L(l,J_3)$ can be obtained from L by deleting the two identical cells $(i_1;j_1)$ and (k,J_3).

In the two cases we have

$$C_{L(l,J_3)} = -\,[C_{L(k,J_1)} - D_{lk}(J_1)] + D_{kl}(J_3)$$

and by Theorem 2 we have

$$C_{L(l,J_3)} = -\,[C_{L(k,J_2)} - D_{lk}(J_2)] + D_{kl}(J_3).$$

THEOREM 4: If $D_{lk}(J_2) > D_{lk}(J_3)$ then either $L(k,J_2)$ or $L(k,J_3)$ is an improving simplex loop.

PROOF: Since $D_{lk}(J_3) = -\,D_{kl}(J_3)$, it follows from Theorem 3 that

$$C_{L(l,J_3)} + C_{L(k,J_2)} = D_{lk}(J_2) - D_{lk}(J_3) > 0$$

$(D_{lk}(J_2) > D_{lk}(J_3))$ and either $C_{L(l,J_3)}$ or $C_{L(k,J_2)}$ is a positive number, e.g., either $L(l,J_3)$ or $L(k,J_2)$ is an improving simplex loop (Definition 2).

COROLLARY: At optimum we have $D_{lk}(J_2) < D_{lk}(J_3)$.

DEFINITION 4: Define J_{lk} by

$$D_{lk}(J_{lk}) = \max_{j \in V_l} D_{lk}(j).$$

REMARK 1: We shall suppose that $D_{lk}(j_1) = D_{lk}(j_2)$ if and only if $j_1 = j_2$. Otherwise, a cost perturbed problem with $C_{ij}^{\epsilon} = C_{ij} + \epsilon^{mi+j}$ can be considered and

$$D_{lk}^{\epsilon}(j_1) - D_{lk}^{\epsilon}(j_2) = C_{lj_1} - C_{kj_1} + \epsilon^{ml+j_1} - \epsilon^{mk+j_1} - C_{lj_2} +$$

$+ C_{kj_2} - \epsilon^{ml+j_2} + \epsilon^{mk+j_2}$ which for sufficiently small $\epsilon > 0$ is equal 0 only for $j_1 = j_2$.

THEOREM 5: If at the optimality $V_{lk} = \phi$ then $D_{lk}(J_{lk}) < D_{lk}(J_{kl})$ $(J_{lk} \neq J_{kl})$, else $D_{lk}(J_{lk}) = D_{lk}(J_{kl})$ $(J_{lk} = J_{kl})$.

PROOF: If $V_{lk} = \phi$ then $D_{lk}(J_{lk}) \neq D_{lk}(J_{kl})$ otherwise, (by cost-perturbation) $J_{lk} = J_{kl}$ and $V_{lk} \neq \phi$.

The first part of Theorem 5 follows now immediately from the corollary of Theorem 4.

If $V_{lk} \neq \phi$ and j is the unique element of V_{lk} then by the definition of J_{lk} and from $j \in V_l$ we have $D_{lk}(j) \leqslant D_{lk}(J_{lk})$.

Let us show that $j = J_{lk}$. Suppose that $j \neq J_{lk}$ then we have $D_{lk}(j) < D_{lk}(J_{lk})$, (Definition 4) and the simplex loop

$L(k,J_{lk}) = (k,j)\ (l,j)\ (l,J_{lk})\ (k,J_{lk})$ will be an improving simplex loop since

$$C_{L(k,J_{lk})} = D_{kl}(j) + D_{lk}(J_{lk}) = D_{lk}(J_{lk}) - D_{lk}(j) > 0$$

contradicting the fact that we have optimality.

Thus, $j = J_{lk}$. By the same argument we have $j = J_{kl}$.

A simple algorithm consists of

1) Computing the differences $D_{lk}\ (J_{lk})$,

2) Comparing $D_{lk}(J_{lk})$ to $D_{lk}(J_{kl})$.

If $D_{lk}(J_{lk}) > D_{lk}(J_{kl})$ (or if $J_{lk} \neq J_{kl}$ for non-empty V_{lk}) we improve our solution, using all the nonbasic cells (l,J) or (k,J) where $J \in (V_l U V_k)$ such that $D_{lk}(J_{lk}) < D_{lk}(J) < D_{lk}(J_{kl})$ by searching only the first loop involving the rows l and k.

The other loops will be obtained by changing the last two cells keeping the $2k-2$ first cells in the same order or in the opposite order (Theorem 2 and Theorem 3).

REMARK: In order to assure that the first loop will not be a shortened loop, this loop will be obtained by using a fundamental artificial destination J with only one basic cell in the k row with $x_{k,j} = \epsilon$.

The proposed technique was applied to each pair of rows (l,k) until $D_{lk}(J_{lk}) \leqslant D_{lk}(J_{kl})$ for all $1 \leqslant l \neq k \leqslant m$. At that point the MODI method was implemented. Performing a MODI iteration frequently caused $D_{lk}(J_{lk}) > D_{lk}(J_{kl})$ for some $1 \leqslant l \neq k \leqslant m$ which would enable further utilization of the proposed technique. However, for the purpose of the present experiment the proposed technique was not reactivated after the initial processing. (See Table 1)

The storage and time requirements of the lists J_{lk} when updated at each MODI iteration are fully discussed in [1].

One possible way to update this list may be described as follows: For each l the destinations of $j \in V_l$ are ordered in $m-1$ sequences P_{lk} $(1 \leqslant k \neq l \leqslant m)$ of increasing $D_{lk}(j)$. Thus $P_{lk} = \{j_1; j_2 \ldots; j_{N_l}\}$ (N_l — the number of elements in V_l) $D_{lk}(j_1) < D_{lk}(j_2) \ldots < D_{lk}(j_{N_l})$ (equality excluded because of the supposed cost-perturbation). These P_{lk} sequences are organized in heaps. Adding or deleting an item from a heap requires $0(\log N_l) < 0(\log n)$ computer operations. Since at each simplex iteration only one basic cell, say (σ, τ), becomes nonbasic and one nonbasic cell, say (s,t), becomes basic, we have to update $2(m-1)$ heaps ($P_{\sigma\rho}$ and P_{sr} for all $\rho \neq \sigma$, $r \neq s$), which amounts to $0(m \log n)$ computer operations per simplex iteration. (heaps, see [2]).

REFERENCES

[1] Brandt, A. and J. Intrator, "Fast Algorithms for Long Transportation Problems," Computers and Operations Research *5*, 263-271 (1978).

[2] Knuth, D.E., *The Art of Computer Programming, 3, Sorting and Searching*, Addison Wesley (1973).

ON THE GENERATION OF DEEP DISJUNCTIVE CUTTING PLANES*

Hanif D. Sherali and C. M. Shetty

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

In this paper we address the question of deriving deep cuts for nonconvex disjunctive programs. These problems include logical constraints which restrict the variables to at least one of a finite number of constraint sets. Based on the works of Balas, Glover, and Jeroslow, we examine the set of valid inequalities or cuts which one may derive in this context, and defining reasonable criteria to measure depth of a cut we demonstrate how one may obtain the "deepest" cut. The analysis covers the case where each constraint set in the logical statement has only one constraint and is also extended for the case where each of these constraint sets may have more than one constraint.

1. INTRODUCTION

A Disjunctive Program is an optimization problem where the constraints represent logical conditions. In this study we are concerned with such conditions expressed as linear constraints. Several well-known problems can be posed as disjunctive programs, including the zero-one integer programs. The logical conditions may include conjunctive statements, disjunctive statements, negation and implication as discussed in detail by Balas [1,2]. However, an implication can be restated as a disjunction, and conjunctions and negations lead to a polyhedral constraint set. Thus, this study deals with the harder problem involving disjunctive restrictions which are essentially nonconvex problems.

It is interesting to note that disjunctive programming provides a powerful unifying theory for cutting plane methodologies. The approach taken by Balas [2] and Jeroslow [14] is to characterize all valid cutting planes for disjunctive programs. As such, it naturally leads to a statement which subsumes prior efforts at presenting an unified theory using convex sets, polar sets and level sets of gauge functions [1,2,5,6,8,13,14]. On the other hand, the approach taken by Glover [10] is to characterize all valid cutting planes through relaxations of the original disjunctive program. Constraints are added sequentially, and when all the constraints are considered Glover's, result is equivalent to that of Balas and Jeroslow. Glover's approach is a constructive procedure for generating valid cuts, and may prove useful algorithmically.

The principal thrust of the methodologies of disjunctive programming is the generation of cutting planes based on the linear logical disjunctive conditions in order to solve the corresponding nonconvex problem. Such methods have been discussed severally by Balas [1,2,3], Glover [8], Glover, Klingman and Stutz [11], Jeroslow [14] and briefly by Owen [17]. But the most fundamental and important result of disjunctive programming has been stated by

*This paper is based upon work supported by the National Science Foundation under Grant No. ENG-77-23683.

Balas [1,2] and Jeroslow [14], and in a different context by Glover [10]. It unifies and subsumes several earlier statements made by other authors and is restated below. This result not only provides a basis for unifying cutting plane theory, but also provides a different perspective for examining this theory. In order to state this result, we will need to use the following notation and terminology.

Consider the linear inequality systems S_h, $h \in H$ given by

$$S_h = \{x: A^h x \geqslant b^h,\ x \geqslant 0\},\ h \in H \tag{1.1}$$

where H is an appropriate index set. We may state a disjunction in terms of the sets S_h, $h \in H$ as a condition which asserts that a feasible point must satisfy at least one of the constraint S_h, $h \in H$. Notationally, we imply by such a disjunction, the restriction $x \in \bigcup_{h \in H} S_h$. Based on this disjunction, an inequality $\pi^t x \geqslant \pi_0$ will be considered a *valid inequality* or a *valid disjunctive cut* if it is satisfied for each $x \in \bigcup_{h \in H} S_h$. (The superscript t will throughout be taken to denote the transpose operation). Finally, for a set of vectors $\{v^h: h \in H\}$, where $v^h = (v_1^h, \ldots, v_n^h)$ for each $h \in H$, we will denote by $\sup_{h \in H} (v^h)$, the pointwise supremum $v = (v_1, \ldots, v_n)$ of the vectors v^h, $h \in H$, such that $v_j = \sup_{h \in H} \{v_j^h\}$ for $j = 1, \ldots, n$.

Before proceeding, we note that a condition which asserts that a feasible point must satisfy at least p of some q sets, $p \leqslant q$, may be easily transformed into the above disjunctive statement by letting each S_h denote the conjunction of the q original sets taken p at a time. Thus, $H = \left\{1, \ldots, \binom{q}{p}\right\}$ in this case. Now consider that following result.

THEOREM 1: (Basic Disjunctive Cut Principle) — Balas [1,2], Glover [10], Jeroslow [14]

Suppose that we are given the linear inequality systems S_h, $h \in H$ of Equation (1.1), where $|H|$ may or may not be finite. Further, suppose that a feasible point must satisfy at least one of these systems. Then, for any choice of nonnegative vectors λ^h, $h \in H$, the inequality

$$\left(\sup_{h \in H} (\lambda^h)^t A^h\right) x \geqslant \inf_{h \in H} (\lambda^h)^t b^h \tag{1.2}$$

is a valid disjunctive cut. Furthermore, if every system S_h, $h \in H$ is consistent, and if $|H| < \infty$, then for any valid inequality $\sum_{j=1}^{n} \pi_j x_j \geqslant \pi_0$, there exist nonnegative vectors λ^h, $h \in H$ such that $\pi_0 \leqslant \inf_{h \in H} (\lambda^h)^t b^h$ and for $j = 1, \ldots, n$, the jth component of $\sup (\lambda^h)^t A^h$ does not exceed π_j.

The forward part of the above theorem was originally proved by Balas [2] and the converse part by Jeroslow [14]. This theorem has also been independently proved by Glover [10] in a somewhat different setting. The theorem merely states that given a disjunction $x \in \bigcup_{h \in H} S_h$, one may generate a valid cut (1.2) by specifying any nonnegative values for the vectors λ^h, $h \in H$. The versatility of the latter choice is apparent from the converse which asserts that so long as we can identify and delete any inconsistent systems, S_h, $h \in H$, then given any valid cut $\pi^t x \geqslant \pi_0$, we may generate a cut of the type (1.2) by suitably selecting values for the parameters λ^h, $h \in H$ such that for any x belonging to the nonnegative orthant of R^n, if (1.2) holds then we must have $\pi^t x \geqslant \pi_0$. In other words, we can make a cut of the type (1.2) uniformly dominate any given valid inequality or cut. Thus, any valid inequality is either a special case of

(1.2) or may be strictly dominated by a cut of type (1.2). In this connection, we draw the reader's attention to the work of Balas [1] in which several convexity/intersection cuts discussed in the literature are recovered from the fundamental disjunctive cut. Note that since the inequality (1.2) defines a closed convex set, then for it to be valid, it must necessarily contain the polyhedral set

$$S = \text{convex hull of } \bigcup_{h \in H} S_h. \tag{1.3}$$

Hence, one may deduce that a desirable deep cut would be a facet of S, or at least would support it. Indeed, Balas [3] has shown how one may generate with some difficulty cuts which contain as a subset, the facets of S when $|H| < \infty$. Our approach to developing deep disjunctive cuts will bear directly on Theorem 1. Specifically, we will be indicating how one may specify values for parameters λ^h to provide supports of S, and will discuss some specific criteria for choosing among supports. We will be devoting our attention to the following two disjunctions titled DC1 and DC2. We remark that most disjunctive statements can be cast in the format of DC2. Disjunction DC1 is a special case of disjunction DC2, and is discussed first because it facilitates our presentation.

DC1:

Suppose that each systems S_h is comprised of a single linear inequality, that is, let

$$S_h = \left\{ x : \sum_{j=1}^{n} a_{1j}^h x_j \geqslant b_1^h,\ x \geqslant 0 \right\} \text{ for } h \in H = \{1, \ldots, \hat{h}\} \tag{1.4}$$

where we assume that $\hat{h} = |H| < \infty$ and that each inequality in S_h, $h \in H$ is stated with the origin as the current point at which the disjunctive cut is being generated. Then, the disjunctive statement DC1 is that at least one of the sets S_h, $h \in H$ must be satisfied. Since the current point (origin) does not satisfy this disjunction, we must have $b_1^h > 0$ for each $h \in H$. Further, we will assume, without loss of generality, that for each $h \in H$, $a_{1j}^h > 0$ for some $j \in \{1, \ldots, n\}$ or else, S_h is inconsistent and we may disregard it.

DC2:

Suppose each system S_h is comprised of a set of linear inequalities, that is, let

$$S_h = \left\{ x : \sum_{j=1}^{n} a_{ij}^h x_j \geqslant b_i^h \text{ for each } i \in Q_h,\ x \geqslant 0 \right\} \text{ for } h \in H = \{1, \ldots, \hat{h}\} \tag{1.5}$$

where Q_h, $h \in H$ are appropriate constraint index sets. Again, we assume that $\hat{h} = |H| < \infty$ and that the representation in (1.5) is with respect to the current point as the origin. Then, the disjunctive statement DC2 is that at least one of the sets S_h, $h \in H$ must be satisfied. Although it is not necessary here for $b_i^h > 0$ for all $i \in Q_h$ one may still state a valid disjunction by deleting all constraints with $b_i^h \leqslant 0$, $i \in Q_h$ from each set S_h, $h \in H$. Clearly a valid cut for the relaxed constraint set is valid for the original constraint set. We will thus obtain a cut which possibly is not as strong as may be derived from the original constraints. To aid in our development, we will therefore assume henceforth that $b_i^h > 0$, $i \in Q_h$, $h \in H$.

Before proceeding with our analysis, let us briefly comment on the need for deep cuts. Although intuitively desirable, it is not always necessary to seek a deepest cut. For example, if one is using cutting planes to implicitly search a feasible region of discrete points, then all cuts which delete the same subset of this discrete region may be equally attractive irrespective of their depth relative to the convex hull of this discrete region. Such a situation arises, for example, in the work of Majthay and Whinston [16]. On the other hand, if one is confronted with

the problem of iteratively exhausting a feasible region which is not finite, as in [20] for example, then indeed deep cuts are meaningful and desirable.

2. DEFINING SUITABLE CRITERIA FOR EVALUATING THE DEPTH OF A CUT

In this section, we will lay the foundation for the concepts we propose to use in deriving deep cuts. Specifically, we will explore the following two criteria for deriving a deep cut:

(i) Maximize the euclidean distance between the origin and the nonnegative region feasible to the cutting plane

(ii) Maximize the rectilinear distance between the origin and the nonnegative region feasible to the cutting plane.

Let us briefly discuss the choice of these criteria. Referring to Figure 1(a) and (b), one may observe that simply attempting to maximize the euclidean distance from the origin to the cut can favor weaker over strictly stronger cuts. However, since one is only interested in the subset of the nonnegative orthant feasible to the cuts, the choice of criterion (i) above avoids such anamolies. Of course, as Figure 1(b) indicates, it is possible for this criterion to be unable to recognize dominance, and treat two cuts as alternative optimal cuts even through one cut dominates the other.

Let us now proceed to characterize the euclidean distance from the origin to the nonnegative region feasible to a cut

$$\sum_{j=1}^{n} z_j x_j \geqslant z_0, \text{ where } z_0 > 0,\ z_j > 0 \text{ for some } j \in \{1, \ldots, n\}. \tag{2.1}$$

The required distance is clearly given by

$$\theta_e = \text{minimum } \{||x||: \sum_{j=1}^{n} z_j x_j \geqslant z_0,\ x \geqslant 0\}. \tag{2.2}$$

Consider the following result.

LEMMA 1: Let θ_e be defined by Equations (2.1) and (2.2). Then

$$\theta_e = \frac{z_0}{||y||} \tag{2.3}$$

where,

$$y = (y_1, \ldots, y_n),\ y_j = \text{maximum } \{0, z_j\},\ j = 1, \ldots, n. \tag{2.4}$$

PROOF: Note that the solution $x^* = \left(\frac{z_0}{||y||^2}\right) y$ is feasible to the problem in (2.2) with $||x^*|| : \frac{z_0}{||y||}$. Moreover, for any x feasible to (2.2), we have, $z_0 \leqslant \sum_{j=1}^{n} z_j x_j \leqslant \sum_{j=1}^{n} y_j x_j \leqslant ||y||\ ||x||$, or that, $||x|| \geqslant \frac{z_0}{||y||}$. This completes the proof.

Now, let us consider the second criterion. The motivation for this criterion is similar to that for the first criterion and moreover, as we shall see below, the use of this criterion has

intuitive appeal. First of all, given a cut (2.1), let us characterize the rectilinear distance from the origin to the nonnegative region feasible to this cut. This distance is given by

$$\theta_r = \text{minimum } \{|x|: \sum_{j=1}^{n} z_j x_j \geqslant z_0,\ x \geqslant 0\}, \text{ when } |x| \sum_{j=1}^{n} x_j. \tag{2.5}$$

Consider the following result.

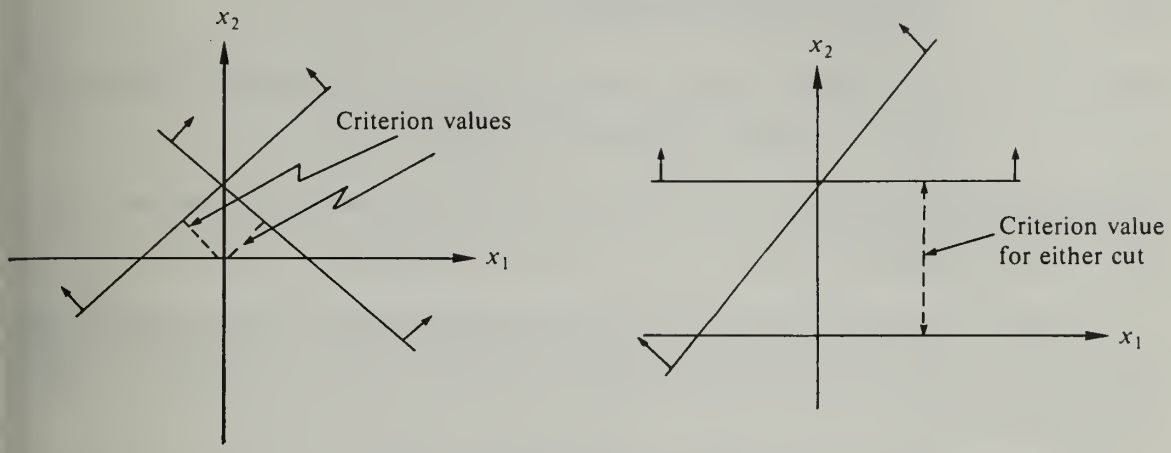


FIGURE 1. Recognition of dominance

LEMMA 2: Let θ_r be defined by Equations (2.1) and (2.5). Then,

$$\theta_r = \frac{z_0}{z_m} \text{ where } z_m = \underset{j=1,\ldots,n}{\text{maximum}}\ z_j. \tag{2.6}$$

PROOF: Note that the solution $x^* = (0, \ldots, \frac{z_0}{z_m}, \ldots 0)$, with the m th component being non-zero, is feasible to the problem in (2.5) with $|x^*| = \frac{z_0}{z_m}$. Moreover, for any x feasible to (2.5), we have,

$$\frac{z_0}{z_m} \leqslant \sum_{j=1}^{n} \frac{z_j}{z_m} x_j \leqslant \sum_{j=1}^{n} x_j = |x|.$$

This completes the proof.

Note from Equation (2.6) that the objective of maximizing θ_r is equivalent to finding a cut which maximizes the smallest positive intercept made on any axis. Hence, the intuitive appeal of this criterion.

3. DERIVING DEEP CUTS FOR DC1

It is very encouraging to note that for the disjunction DC1 we are able to derive a cut which not only simultaneously satisfies both the criterion of Section 2, but which is also a facet of the set S of Equation (1.3). This is a powerful statement since all valid inequalities are given through (1.2) and none of these can strictly dominate a facet of S.

We will find it more convenient to state our results if we normalize the linear inequalities (1.4) by dividing through by their respective, positive, right-hand-sides. Hence, let us assume without loss of generality that

$$S_h = \left\{ x : \sum_{j=1}^{n} a_{1j}^h x_j \geq 1,\ x \geq 0 \right\} \text{ for } h \in H = \{1, \ldots, \hat{h}\}. \tag{3.1}$$

Then the application of Theorem 1 to the disjunction DC1 yields valid cuts of the form:

$$\sum_{j=1}^{n} \left\{ \max_{h \in H} \lambda_1^h a_{1j}^h \right\} x_j \geq \min_{h \in H} \{\lambda_1^h\} \tag{3.2}$$

where λ_1^h, $h \in H$ are nonnegative scalars. Again, there is no loss of generality in assuming that

$$\sum_{h \in H} \lambda_1^h = 1,\ \lambda_1^h \geq 0,\ h \in H = \{1, \ldots, \hat{h}\} \tag{3.3}$$

since we will not allow all λ_1^h, $h \in H$ to be zero. This is equivalent to normalizing (3.2) by dividing through by $\sum_{h \in H} \lambda_1^h$.

Theorem 2 below derives two cuts of the type (3.2), both of which simultaneously achieve the two criteria of the foregoing section. However, the second cut uniformly dominates the first cut. In fact, no cut can strictly dominate the second cut since it is shown to be a facet of S defined by (1.3).

THEOREM 2: Consider the disjunctive statement DC1 where S_h is defined by (3.1) and is assumed to be consistent for each $h \in H$. Then the following results hold:

(a) Both the criteria of Section 2 are satisfied by letting $\lambda_1^h = \lambda_1^{h^*}$ where

$$\lambda_1^{h^*} = 1/\hat{h} \qquad \text{for } h \in H \tag{3.4}$$

in inequality (3.2) to obtain the cut

$$\sum_{j=1}^{n} a_{1j}^{*} x_j \geq 1, \text{ where } a_{1j}^{*} = \max_{h \in H} a_{1j}^h, \text{ for } j = 1, \ldots, n. \tag{3.5}$$

(b) Further, defining

$$\gamma_1^h = \underset{j : a_{1j}^h > 0}{\text{minimum}} \{a_{1j}^{*}/a_{1j}^h\} > 0,\ h \in H \tag{3.6}$$

and letting $\lambda_1^h = \lambda_1^{h^{**}}$, where

$$\lambda_1^{h^{**}} = \gamma_1^h / \sum_{p \in H} \gamma_1^p \qquad \text{for } h \in H \tag{3.7}$$

in inequality (3.2), we obtain a cut of the form

$$\sum_{j=1}^{n} a_{1j}^{**} x_j \geq 1, \text{ where } a_{1j}^{**} = \max_{h \in H} a_{1j}^h \gamma_1^h \text{ for } j = 1, \ldots, n \tag{3.8}$$

which again satisfies both the criteria of Section 2.

(c) The cut (3.8) uniformly dominates the cut (3.5); in fact,

$$a_{1j}^{**} \begin{cases} = a_{1j}^{*} \text{ if } a_{1j}^{*} > 0 \\ \leq a_{1j}^{*} \text{ if } a_{1j}^{*} \leq 0 \end{cases}, \quad j = 1, \ldots, n. \tag{3.9}$$

(d) The cut (3.8) is a facet of the set S of Equation (1.3).

PROOF:

(a) Clearly, $\lambda_1^h = 1/\hat{h}$, $h \in H$ leads to the cut (3.5) from (3.2). Now consider the euclidean distance criterion of maximizing θ_e (*or* θ_e^2) of Equation (2.3). For cut (3.5), the value of θ_e^2 is given by

$$(\theta_e^*)^2 = 1\Big/\sum_{j=1}^{n} (y_j^*)^2 > 0 \text{ where } y_j^* = \max\{0, a_{1j}^*\},\ j = 1, \ldots, n. \tag{3.10}$$

Now, for any choice λ_1^h, $h \in H$,

$$\theta_e^2 = \Big[\min_{h\in H}(\lambda_1^h)\Big]^2\Big/\sum_{j=1}^{n} y_j^2 = (\lambda_1^p)^2\Big/\sum_{j=1}^{n} y_j^2, \text{ say,} \tag{3.11}$$

where $y_j = \max\{0, \max_{h\in H} \lambda_1^h a_{1j}^h\}$. If $\lambda_1^p = 0$, then $\theta_e = 0$ and noting (3.10), such a choice of parameters λ_1^h, $h \in H$ is suboptimal. Hence, $\lambda_1^p > 0$, whence (3.11) becomes $\theta_e^2 = 1\Big/\sum_{j=1}^{n} \left(\frac{y_j}{\lambda_1^p}\right)^2$. But since $(\lambda_1^h/\lambda_1^p) \geqslant 1$ for each $h \in H$, we get

$$y_j/\lambda_1^p = \max\left\{0, \max_{h\in H}\left\{\left(\frac{\lambda_1^h}{\lambda_1^p}\right) a_{1j}^h\right\}\right\} \geqslant \max\left\{0, \max_{h\in H} a_{1j}^h\right\} = y_j^*.$$

Thus $\theta_e^2 \leqslant (\theta_e^*)^2$ so that the first criterion is satisfied.

Now consider the maximization of θ_r of Equation (2.5), or equivalently Equation (2.6). For the choice (3.4), the value of θ_r is given by

$$\theta_r^* = \frac{1}{\max_j a_{1j}^*} > 0. \tag{3.12}$$

Now, for any choice λ_1^h, $h \in H$, from Equations (2.6), (3.2) we get

$$\theta_r = \left(\min_{h\in H} \lambda_1^h\right)\Big/\left(\max_j \max_{h\in H} \lambda_1^h a_{1j}^h\right) = \lambda_1^p\Big/\max_j \max_{h\in H} \lambda_1^h a_{1j}^h, \text{ say.}$$

As before, $\lambda_1^p = 0$ implies a value of θ_r inferior to θ_r^*. Thus, assume $\lambda_1^p > 0$. Then, $\theta_r = 1\Big/\max_j \max_{h\in H}\left(\frac{\lambda_1^h}{\lambda_1^p}\right) a_{1j}^h$. But $(\lambda_1^h/\lambda_1^p) \geqslant 1$ for each $h \in H$ and in evaluating θ_r, we are interested only in those $j \in \{1, \ldots, n\}$ for which $a_{1j}^h > 0$ for some $h \in H$. Thus, $\theta_r \leqslant 1/\max_j \max_{h\in H} a_{1j}^h = \theta_r^*$, so that the second criterion is also satisfied. This proves part (a).

(b) and (c). First of all, let us consider the values taken by γ_1^h, $h \in H$. Note from the assumption of consistency that γ_1^h, $h \in H$ are well defined. From (3.5), (3.6), we must have $\gamma_1^h \geqslant 1$ for each $h \in H$. Moreover, if we define from (3.5)

$$H^* = \{h \in H: a_{1k}^h = a_{1k}^* > 0 \text{ for some } k \in \{1, \ldots, n\}\} \tag{3.13}$$

then clearly $H^* \neq \{\phi\}$ and for $h \in H^*$, Equation (3.6) implies $\gamma_1^h \leqslant 1$. Thus,

$$\gamma_1^h \begin{cases} = 1 & \text{for } h \in H^* \\ > 1 & \text{for } h \notin H^*. \end{cases} \tag{3.14}$$

Hence,

$$\min_{h \in H} \gamma_1^h = 1 \tag{3.15}$$

or that, using (3.7) in (3.2) yields a cut of the type (3.8), where,

$$a_{1j}^{\bullet\bullet} = \max_{h \in H} a_{1j}^h \gamma_1^h, \; j = 1, \ldots, n. \tag{3.16}$$

Now, let us establish relationship (3.9). Note from (3.5) that if $a_{1j}^{\bullet} \leqslant 0$, then $a_{1j}^h \leqslant 0$ for each $h \in H$ and hence, using (3.14), (3.16), we get that (3.9) holds. Next, consider $a_{1j}^{\bullet} > 0$ for some $j \in \{1, \ldots, n\}$. From (3.13), (3.14), (3.16), we get

$$a_{1j}^{\bullet\bullet} = \max\{\max_{h \in H} a_{1j}^h, \; \max_{\substack{h \in H^* \\ a_{1j}^h > 0}} a_{1j}^h \gamma_1^h\} \tag{3.17}$$

where we have not considered $h \notin H^*$ with $a_{1j}^h \leqslant 0$ since $a_{1j}^{\bullet\bullet} > 0$. But for $h \notin H^*$ with $a_{1j}^h > 0$, we get from (3.5), (3.6)

$$a_{1j}^h \gamma_1^h = a_{1j}^h \left[\min_{k : a_{1k}^h > 0} \left\{ \frac{\max_{r \in H} a_{1k}^r}{a_{1k}^h} \right\} \right] \leqslant a_{1j}^h \left\{ \frac{\max_{r \in H} a_{1j}^r}{a_{1j}^h} \right\} = \max_{r \in H} a_{1j}^r. \tag{3.18}$$

Using (3.18) in (3.17) yields $a_{1j}^{\bullet\bullet} = a_{1j}^{\bullet}$, which establishes (3.9).

Finally, we show that (3.8) satisfies both the criteria of Section 2. This part follows immediately from (3.9) by noting that the cut (3.5) yields $\theta_e = \theta_e^{\bullet}$ of (3.10) and $\theta_r = \theta_r^{\bullet}$ of (3.12). This completes the proofs of parts (b) and (c).

(d) Note that since (3.8) is valid, any $x \in S$ satisfies (3.8). Hence, in order to show that (3.8) defines a facet of S, it is sufficient to identify n affinely independent points of S which satisfy (3.8) as an equality, since clearly, dim $S = n$. Define

$$J_1 = \{j \in \{1, \ldots, n\} : a_{1j}^{\bullet\bullet} > 0\} \text{ and let } J_2 = \{1, \ldots, n\} - J_1. \tag{3.19}$$

Consider any $p \in J_1$, and let

$$e_p = (0, \ldots, \frac{1}{a_{1p}^{\bullet\bullet}}, \ldots, 0), \; p \in J_1 \tag{3.20}$$

have the non-zero term in the p^{th} position. Now, since $p \in J_1$, (3.9) yields

$$a_{1p}^{\bullet\bullet} = a_{1p}^{\bullet} = \max_{h \in H} a_{1p}^h = a_{1p}^{h_p}, \text{ say,}$$

Hence, $e_p \in S_{h_p}$ and so, $e_p \in S$ and moreover, e_p satisfies (3.8) as an equality. Thus, e_p, $p \in J_1$ qualify as $|J_1|$ of the n affinely independent points we are seeking.

Now consider a $q \in J_2$. Let us show that there exists an S_{h_q} satisfying

$$\gamma_1^{h_q} a_{1p}^{h_q} = a_{1p}^{\bullet\bullet} \text{ for some } p \in J_1$$

and

$$\gamma_1^{h_q} a_{1q}^{h_q} = a_{1q}^{\bullet\bullet}. \tag{3.21}$$

From Equation (3.16), we get $a_{1q}^{\bullet\bullet} = \max_{h \in H} a_{1q}^{h} \gamma_1^{h} = a_{1q}^{h_q} \gamma_1^{h_q}$, say. Then for this $h_q \in H$, Equation (3.6) yields $\gamma_1^{h_q} = \underset{j: a_{1j}^{h_q} > 0}{\text{minimum}} \{a_{1j}^{\bullet}/a_{1j}^{h_q}\} = a_{1p}^{\bullet}/a_{1p}^{h_q}$, say. Or, using (3.9), $\gamma_1^{h_q} a_{1p}^{h_q} = a_{1p}^{\bullet} = a_{1p}^{\bullet\bullet} > 0$. Thus (3.21) holds. For convenience, let us rewrite the set S_{h_q} below as

$$S_{h_q} = \{x: a_{1p}^{h_q} x_p + a_{1q}^{h_q} x_q + \sum_{j \neq p,q} a_{1j}^{h_q} x_j \geqslant 1,\ x \geqslant 0\}. \tag{3.22}$$

Now, consider the direction

$$d_q = \begin{cases} (0, \ldots, \dfrac{1}{a_{1p}^{\bullet\bullet}}, \ldots, -\dfrac{1}{a_{1q}^{\bullet\bullet}}, \ldots, 0) & \text{if } a_{1q}^{\bullet\bullet} < 0 \\ (0, \ldots, 0, \ldots, \Delta, \ldots, 0) & \text{if } a_{1q}^{\bullet\bullet} = 0 \end{cases} \tag{3.23}$$

where $\Delta > 0$. Let us show that d_q is a direction for S_{h_q}. Clearly, if $a_{1q}^{\bullet\bullet} = 0$, then from (3.21) $a_{1q}^{h_q} = 0$ and thus (3.22) establishes (3.23). Further, if $a_{1q}^{\bullet\bullet} < 0$ then one may easily verify from (3.21), (3.22), (3.23) that

$$\hat{e}_p = (0, \ldots, \gamma_1^{h_q}/a_{1p}^{\bullet\bullet}, \ldots, 0) \in S_{h_q} \text{ and } \hat{e}_p + \delta[\gamma_1^{h_q} d_q] \in S_{h_q} \text{ for each } \delta \geqslant 0$$

where $\hat{e}_p$ has the non-zero term at position p. Thus, d_q is a direction for S_{h_q}. It can be easily shown that this implies d_q is a direction for S. Since $e_p = (0, \ldots, \dfrac{1}{a_{1p}^{\bullet\bullet}}, \ldots, 0)$ of Equation (3.20) belongs to S, then so does $(e_p + d_q)$. But $(e_p + d_q)$ clearly satisfies (3.8) as an equality. Hence, we have identified n points of S, which satisfy the cut (3.8) as an equality, of the type

$$\left.\begin{aligned} e_p &= (0, \ldots, \frac{1}{a_{1p}^{\bullet\bullet}}, \ldots, 0) \quad \text{for } p \in J_1 \\ e_q &= d_q + e_p \text{ for some } p \in J_1, \text{ for each } q \in J_2 \end{aligned}\right\} \tag{3.24}$$

where d_q is given by (3.23). Since these n points are clearly affinely independent, this completes the proof.

It is interesting to note that the cut (3.5) has been derived by Balas [2] and by Glover [9, Theorem 1]. Further, the cut (3.8) is precisely the strengthened negative edge extension cut of Glover [9, Theorem 2]. The effect of replacing $\lambda_1^{h^*}$ defined in (3.4) by $\lambda_1^{h^{**}}$ defined in (3.7) is equivalent to the translation of certain hyperplanes in Glover's theorem. We have hence shown through Theorem 2 how the latter cut may be derived in the context of disjunctive programming, and be shown to be a facet of the convex hull of feasible points. Further, both (3.5) and (3.8) have been shown to be alternative optima to the two criteria of Section 2.

In generalizing this to disjunction DC2, we find that such an ideal situation no longer exists. Nevertheless, we are able to obtain some useful results. But before proceeding to DC2, let us illustrate the above concepts through an example.

EXAMPLE: Let $H = \{1,2\}$, $n = 3$ and let DC1 be formulated through the sets

$$S_1 = \{x: x_1 + 2x_2 - 4x_3 \geq 1,\ x \geq 0\}, S_2 = \{x: \frac{x_1}{2} + \frac{x_2}{3} - 2x_3 \geq 1,\ x \geq 0\}.$$

The cut (3.5), i.e., $\Sigma a^{\bullet}_{1j} x_j \geq 1$, is $x_1 + 2x_2 - 2x_3 \geq 1$. From (3.6),

$$\gamma_1^1 = \min\left\{\frac{1}{1}, \frac{2}{2}\right\} = 1 \text{ and } \gamma_1^2 = \min\left\{\frac{1}{1/2}, \frac{2}{1/3}\right\} = 2.$$

Thus, through (3.7), or more directly, from (3.16), the cut (3.8), i.e., $\Sigma\, a^{\bullet\bullet}_{1j} x_j \geq 1$ is $x_1 + 2x_2 - 4x_3 \geq 1$. This cut strictly dominates the cut (3.5) in this example, though both have the same values $1/\sqrt{5}$ and $1/2$ respectively for θ_e and θ_r of Equations (2.2) and (2.5).

4. DERIVING DEEP CUTS FOR DC2

To begin with, let us make the following interesting observation. Suppose that for convenience, we assume without loss of generality as before, that $b_i^h = 1$, $i \in Q_h$, $h \in H$ in Equation (1.4). Thus, for each $h \in H$, we have the constraint set

$$S_h = \left\{x: \sum_{j=1}^{n} a_{ij}^h x_j \geq 1,\ i \in Q_h,\ x \geq 0\right\}. \tag{4.1}$$

Now for each $h \in H$, let us multiply the constraints of S_h by corresponding scalars $\delta_i^h \geq 0$, $i \in Q_h$ and add them up to obtain the surrogate constraint

$$\sum_{j=1}^{n} \left\{\sum_{i \in Q_h} \delta_i^h a_{ij}^h\right\} x_j \geq \sum_{i \in Q_h} \delta_i^h,\ h \in H. \tag{4.2}$$

Further, assuming that not all δ_i^h are zero for $i \in Q_h$, (4.2) may be re-written as

$$\sum_{j=1}^{n} \left\{\sum_{i \in Q_h} \left[\frac{\delta_i^h}{\left(\sum_{p \in Q_h} \delta_p^h\right)}\right] a_{ij}^h\right\} x_j \geq 1,\ h \in H. \tag{4.3}$$

Finally, denoting $\delta_i^h \Big/ \sum_{p \in Q_h} \delta_p^h$ by λ_i^h for $i \in Q_h$, $h \in H$, we may write (4.3) as

$$\sum_{j=1}^{n} \left\{\sum_{i \in Q_h} \lambda_i^h a_{ij}^h\right\} x_j \geq 1 \text{ for each } h \in H \tag{4.4}$$

where,

$$\sum_{i \in Q_h} \lambda_i^h = 1 \text{ for each } h \in H,\ \lambda_i^h \geq 0 \text{ for } i \in Q_h,\ h \in H. \tag{4.5}$$

Observe that by surrogating the constraints of (4.1) using parameters λ_i^h, $i \in Q_h$, $h \in H$ satisfying (4.5), we have essentially represented DC2 as DC1 through (4.4). In other words, since $x \in S_h$ implies x satisfies (4.4) for each $h \in H$, then given λ_i^h, $i \in Q_h$, $h \in H$, DC2 implies that at least one of (4.4) must be satisfied. Now, whereas Theorem 1 would directly employ (4.2) to derive a cut, since we have normalized (4.2) to obtain (4.4), we know from the previous section that the optimal strategy is to derive a cut (3.8) using inequalities (4.4).

Now let us consider in turn the two criteria of Section 2.

4.1. Euclidean Distance-Based Criterion

Consider any selection of values for the parameters λ_i^h, $i \in Q_h$, $h \in H$ satisfying (4.5) and let the corresponding disjunction DC1 derived from DC2 be that at least one of (4.4) must hold. Then, Theorem 2 tells us through Equations (3.5), (3.10) that the euclidean distance criterion value for the resulting cut (3.8) is

$$\theta_e(\lambda) = 1 \Big/ \sqrt{\sum_{j=1}^{n} y_j^2} \tag{4.6}$$

where,

$$y_j = \max\{0, z_j\}, \; j = 1, \ldots, n \tag{4.7}$$

and

$$z_j = \max_{h \in H} \left\{ \sum_{i \in Q_h} \lambda_i^h a_{ij}^h \right\}, \; j = 1, \ldots, n. \tag{4.8}$$

Thus, the criterion of Section 2 seeks to

$$\text{maximize } \{\theta_e(\lambda) : \lambda = (\lambda_i^h) \text{ satisfies (4.5)}\} \tag{4.9}$$

or equivalently, to

$$\text{minimize } \{\sum_{j=1}^{n} y_j^2 : \text{ (4.5), (4.7), (4.8) are satisfied}\}. \tag{4.10}$$

It may be easily verified that the problem of (4.10) may be written as

$$\text{PD}_2: \quad \text{minimize } \sum_{j=1}^{n} y_j^2 \tag{4.11}$$

$$\text{subject to } y_j \geqslant \sum_{i \in Q_h} \lambda_i^h a_{ij}^h \text{ for each } h \in H \text{ for each } j = 1, \ldots, n \tag{4.12}$$

$$\sum_{i \in Q_h} \lambda_i^h = 1 \text{ for each } h \in H \tag{4.13}$$

$$\lambda_i^h \geqslant 0 \; i \in Q_h, \; h \in H \tag{4.14}$$

Note that we have deleted the constraints $y_j \geqslant 0$, $j = 1, \ldots, n$ since for any feasible λ_i^h, $i \in Q_h$, $h \in H$, there exists a dominant solution with nonnegative $y_j = j = 1, \ldots, n$. This relaxation is simply a matter of convenience in our solution strategy.

Before proposing a solution procedure for Problem PD_2, let us make some pertinent remarks. Note that Problem PD_2 has the purpose of generating parameters λ_i^h, $i \in Q_h$, $h \in H$ which are to be used to obtain the surrogate constraints (4.4). Thereafter, the cut that we derive for the disjunction DC2 is the cut (3.8) obtained from the statement that at least one of (4.4) must hold. Hence, Problem PD_2 attempts to find values for λ_i^h, $i \in Q_h$, $h \in H$, such that this resulting cut achieves the euclidean distance criterion.

Problem PD_2 is a convex quadratic program for which the Kuhn-Tucker conditions are both necessary and sufficient. Several efficient simplex-based quadratic programming procedures are available to solve such a problem. However, these procedures require explicit handling of the potentially large number of constraints in Problem PD_2. On the other hand, the

subgradient optimization procedure discussed below takes full advantage of the problem structure. We are first able to write out an almost complete solution to the Kuhn-Tucker system. We will refer to this as a *partial solution*. In case we are unable to either actually construct a complete solution or to assert that a feasible completion exists, then through the construction procedure itself, we have a subgradient direction available. Moreover, this latter direction is very likely to be a direction of ascent. We therefore propose to move in the negative of this direction and if necessary, project back onto the feasible region. These iterative steps are now repeated at this new point.

4.1.1 Kuhn-Tucker Systems for PD_2 and Its Implications

Letting u_j^h, $h \in H$, $j = 1, \ldots, n$ denote the lagrangian multipliers for constraints (4.12), t_h, $h \in H$ those for constraints (4.13), and w_i^h, $i \in Q_h$, $h \in H$ those for constraints (4.14), we may write the Kuhn-Tucker optimality conditions as

$$\sum_{h \in H} u_j^h = 2y_j \quad j = 1, \ldots, n \tag{4.15}$$

$$\sum_{j=1}^{n} u_j^h a_{ij}^h + t_h - w_i^h = 0 \text{ for each } i \in Q_h, \text{ and for each } h \in H \tag{4.16}$$

$$u_j^h \left\{ \sum_{i \in Q_h} \lambda_i^h a_{ij}^h - y_j \right\} = 0 \text{ for each } j = 1, \ldots, n \text{ and each } h \in H \tag{4.17}$$

$$\lambda_i^h w_i^h = 0 \text{ for } i \in Q_h, \ h \in H \tag{4.18}$$

$$w_i^h \geqslant 0 \quad i \in Q_h, \ h \in H \tag{4.19}$$

$$u_j^h \geqslant 0 \ j = 1, \ldots, n, \ h \in H. \tag{4.20}$$

Finally, Equations (4.12), (4.13), (4.14) must also hold. We will now consider the implications of the above conditions. This will enable us to construct at least a partial solution to these conditions, given particular values of λ_i^h, $i \in Q_h$, $h \in H$. First of all, note that Equations (4.7), (4.10) and (4.20) imply that

$$y_j \geqslant 0 \text{ for each } j = 1, \ldots, n \tag{4.21}$$

$$y_j = \max\left\{0, \sum_{i \in Q_h} \lambda_i^h a_{ij}^h, \ h \in H\right\} \text{ for } j = 1, \ldots, n. \tag{4.22}$$

Now, having determined values for y_j, $j = 1, \ldots, n$, let us define the sets

$$H_j = \begin{cases} \{\varnothing\} \text{ if } y_j = 0 \\ \{h \in H : y_j = \sum_{i \in Q_h} \lambda_i^h a_{ij}^h > 0\} \end{cases} \quad \text{for } j = 1, \ldots, n. \tag{4.23}$$

Now, consider the determination of u_j^h, $h \in H$, $j = 1, \ldots, n$. Clearly, Equations (4.15), (4.17) and (4.20) along with the definition (4.23) imply that for each $j = 1, \ldots, n$

$$u_j^h = 0 \text{ for } h \in H/H_j \text{ and that } \sum_{h \in H_j} u_j^h = 2y_j, \ u_j^h \geqslant 0 \text{ for each } h \in H_j. \tag{4.24}$$

Thus, for any $j \in \{1, \ldots, n\}$, if H_j is either empty or a singleton, the corresponding values for u_j^h, $h \in H$ are uniquely determined. Hence, we have a choice in selecting values for u_j^h, $h \in H_j$

only when $|H_j| \geq 2$ for any $j \in \{1, \ldots, n\}$. Next, multiplying (4.16) by λ_i^h and using (4.18), we obtain

$$\sum_{j=1}^{n} \left[u_j^h \sum_{i \in Q_h} \{\lambda_i^h a_{ij}^h\} \right] + t_h \sum_{i \in Q_h} \lambda_i^h = 0 \text{ for each } h \in H. \tag{4.25}$$

Using Equations (4.13), (4.17), this gives us

$$t_h = -\sum_{j=1}^{n} u_j^h y_j \text{ for each } h \in H. \tag{4.26}$$

Finally, Equations (4.16), (4.26) yield

$$w_i^h = \sum_{j=1}^{n} u_j^h [a_{ij}^h - y_j] \text{ for each } i \in Q_h, \ h \in H. \tag{4.27}$$

Notice that once the variables u_j^h, $h \in H$, $j = 1, \ldots, n$ are fixed to satisfy (4.24), all the variables are uniquely determined. We now show that if the variables w_i^h, $i \in Q_h$, $h \in H$ so determined are nonnegative, we then have a Kuhn-Tucker solution. Since the objective function of PD_2 is convex and the constraints are linear, this solution is also optimal.

LEMMA 2: Let a primal feasible set of λ_i^h, $i \in Q_h$, $h \in H$ be given. Determine values for all variables y_j, u_j^h, t_h, w_i^h using Equations (4.22) through (4.27), selecting an arbitrary solution in the case described in Equation (4.24) if $|H_j| \geq 2$. If $w_i^h \geq 0$, $i \in Q_h$, $h \in H$, then λ_i^h, $i \in Q_h$, $h \in H$ solves Problem PD_2.

PROOF: By construction Equations (4.12), through (4.17), and (4.20) clearly hold. Thus, noting that in our problem the Kuhn-Tucker conditions are sufficient for optimality, all we need to show is that if $w = (w_i^h) \geq 0$ then (4.18) holds. But from (4.17) and (4.27) for any $h \in H$, we have,

$$\sum_{i \in Q_h} \lambda_i w_i^h = \sum_{i \in Q_h} \lambda_i^h \left\{ \sum_{j=1}^{n} u_j^h [a_{ij}^h - y_j] \right\} = \sum_{j=1}^{n} \left\{ u_j^h \left[\sum_{i \in Q_h} \lambda_i^h a_{ij}^h - y_j \right] \right\} = 0$$

for each $h \in H$. Thus, $\lambda_i^h \geq 0$, $w_i^h \geq 0$ $i \in Q_h$, $h \in H$ imply that (4.18) holds and the proof is complete.

The reader may note that in Section 4.1.4 we will propose another stronger sufficient condition for a set of variables λ_i^h, $i \in Q_h$, $h \in H$ to be optimal. The development of this condition is based on a subgradient optimization procedure discussed below.

4.1.2 Subgradient Optimization Scheme for Problem PD

For the purpose of this development, let us use (4.22) to rewrite Problem PD_2 as follows. First of all define

$$\Lambda = \{\lambda = (\lambda_i^h): \text{ constraints (4.13) and (4.14) are satisfied}\} \tag{4.28}$$

and let f: $\Lambda \rightarrow R$ be defined by

$$f(\lambda) = \sum_{j=1}^{n} \left[\text{maximum} \left\{ 0, \sum_{i \in Q_h} \lambda_i^h a_{ij}^h, \ h \in H \right\} \right]^2. \tag{4.29}$$

Then, Problem PD_2 may be written as

$$\text{minimize } \{f(\lambda): \lambda \in \Lambda\}.$$

Note that for each $j = 1, \ldots, n$, $g_j(\lambda) = \max \{0, \sum_{i \in Q_h} \lambda_i^h a_{ij}^h, h \in H\}$ is convex and nonnegative. Thus, $[g_j(\lambda)]^2$ is convex and so $f(\lambda) = \sum_{j=1}^{n} [g_j(\lambda)]^2$ is also convex.

The main thrust of the proposed algorithm is as follows. Having a solution $\bar{\lambda}$ at any stage, we will attempt to construct a solution to the Kuhn-Tucker system using Equations (4.15) through (4.20). If we obtain nonnegative values $\bar{w}_i^h$ for the corresponding variables w_i^h, $i \in Q_h$, $h \in H$, then by Lemma 2 above, we terminate. Later in Section 4.1.7, we will also use another sufficient condition to check for termination. If we obtain no indication of optimality, we continue. Theorem 3 below established that in any case, the vector $w = \bar{w}$ constitutes a subgradient of $f(\cdot)$ at the current point $\bar{\lambda}$. Following Poljak [18,19], we hence take a suitable step in the negative subgradient direction and project back onto the feasible region Λ of Equation (4.28). This completes one iteration. Before presenting Theorem 3, consider the following definition.

DEFINITION 1: Let f: $\Lambda \to R$ be a convex function and let $\lambda \in \Lambda \subset R^m$. Then $\xi \in R^m$ is a *subgradient* of $f(\cdot)$ at $\bar{\lambda}$ if

$$f(\lambda) \geq f(\bar{\lambda}) + \xi^t (\lambda - \bar{\lambda}) \text{ for each } \lambda \in \Lambda.$$

THEOREM 3: Let $\bar{\lambda}$ be a given point in Λ defined by (4.28) and let $\bar{w}$ be obtained from Equations (4.22) through (4.27), with an arbitrary selection of a solution to (4.24).

Then, $\bar{w}$ is a subgradient of $f(\cdot)$ at $\bar{\lambda}$, where f:$\Lambda \to R$ is defined in Equation (4.29).

PROOF. Let y and $\bar{y}$ be obtained through Equation (4.22) from $\lambda \in \Lambda$ and $\bar{\lambda} \in \Lambda$ respectively. Hence,

$$f(\lambda) = \sum_{j=1}^{n} y_j^2 \text{ and } f(\bar{\lambda}) = \sum_{j=1}^{n} \bar{y}_j^2.$$

Thus, from Definition 1, we need to show that

$$\sum_{h \in H} \sum_{i \in Q_h} \bar{w}_i^h (\lambda_i^h - \bar{\lambda}_i^h) \leq \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} \bar{y}_j^2. \tag{4.30}$$

Noting from Equations (4.17), (4.27) that $\sum_{h \in H} \sum_{i \in Q_h} \bar{w}_i^h \bar{\lambda}_i^h = 0$, we have,

$$\sum_{h \in H} \sum_{i \in Q_h} \bar{w}_i^j (\lambda_i^h - \bar{\lambda}_i^h) = \sum_{h \in H} \sum_{i \in Q_h} \bar{w}_i^h \lambda_i^h = \sum_{h \in H} \sum_{i \in Q_h} \sum_{j=1}^{n} \bar{u}_j^h \lambda_i^h [a_{ij}^h - \bar{y}_j]$$

$$= \sum_{h \in H} \sum_{j=1}^{n} \bar{u}_j^h \left(\sum_{i \in Q_h} \lambda_i^h a_{ij}^h \right) - \sum_{h \in H} \sum_{j=1}^{n} \left[\bar{u}_j^h \bar{y}_j \sum_{i \in Q_h} \lambda_i^h \right].$$

Using (4.13) and (4.15), this yields

$$\sum_{h \in H} \sum_{i \in Q_h} \bar{w}_i^h (\lambda_i^h - \bar{\lambda}_i^h) = \sum_{h \in H} \sum_{j=1}^{n} \bar{u}_j^h \left(\sum_{i \in Q_h} \lambda_i^h a_{ij}^h \right) - 2 \sum_{j=1}^{n} \bar{y}_j^2.$$

Combining this with (4.30), we need to show that

$$\sum_{j\in H}\sum_{j=1}^{n}\bar{u}_j^h\left(\sum_{i\in Q_h}\lambda_i^h a_{ij}^h\right) \leqslant \sum_{j=1}^{n} y_j^2 + \sum_{j=1}^{n}\bar{y}_j^2. \tag{4.31}$$

But Equations (4.15), (4.20), (4.22) imply that

$$\sum_{h\in H}\sum_{j=1}^{n}\bar{u}_j^h\left(\sum_{i\in Q_h}\lambda_j^h a_{ij}^h\right) \leqslant \sum_{h\in H}\sum_{j=1}^{n}\bar{u}_j^h y_j = 2\sum_{j=1}^{n} y_j\bar{y}_j \leqslant 2||y||\,||\bar{y}|| \leqslant ||y||^2 + ||\bar{y}||^2$$

so that Equation (4.31) holds. This completes the proof.

Although, given $\bar{\lambda} \in \Lambda$, any solution to Equations (4.22) through (4.27) will yield a subgradient of $f(\cdot)$ at the current point $\bar{\lambda}$, we would like to generate, without expending much effort, a subgradient which is hopefully a direction of ascent. Hence, this would accelerate the cut generation process. Later in Section 4.1.6 we describe one such scheme to determine a suitable subgradient direction. For the present moment, let us assume that we have generated a subgradient $\bar{w}$ and have taken a suitable step size $\bar{\theta}$ in the direction $-\bar{w}$ as prescribed by the subgradient optimization scheme of Held, Wolfe, and Crowder [12]. Let

$$\bar{\bar{\lambda}} = \bar{\lambda} - \bar{\theta}\,\bar{w} \tag{4.32}$$

be the new point thus obtained. To complete the iteration, we must now project $\bar{\bar{\lambda}}$ into Λ, that is, we must determine a new $\bar{\lambda}$ according to

$$\bar{\lambda}_{new} \equiv P_\Lambda(\bar{\bar{\lambda}}) = \text{minimum}\,\{||\lambda - \bar{\bar{\lambda}}||: \lambda \in \Lambda\}. \tag{4.33}$$

The method of accomplishing this efficiently is presented in the next subsection.

4.1.3 Projection Scheme

For convenience, let us define the following linear manifold

$$M_h = \left\{\lambda_i^h,\ i\in Q_h: \sum_{i\in Q_h}\lambda_i^h = 1\right\},\ h\in H \tag{4.34}$$

and let $\bar{M}_h$ be the intersection of M_h with the nonnegative orthant, that is,

$$\bar{M}_h = \{\lambda_i^h,\ i\in Q_h: \sum_{i\in Q_h}\lambda_i^h = 1,\ \lambda_i^h \geqslant 0,\ i\in Q_h\}. \tag{4.35}$$

Note from Equation (4.28) that

$$\Lambda = \bar{M}_1 \times \ldots \times \bar{M}_{|H|}. \tag{4.36}$$

Now, given $\bar{\bar{\lambda}}$, we want to project it onto Λ, that is, determine $\bar{\lambda}_{new}$ from Equation (4.33). Towards this end, for any vector $\alpha = (\alpha_i,\ i\in I)$, where I is a suitable index set for the $|I|$ components of α, let $P(\alpha, I)$ denote the following problem:

$$P(\alpha,I): \qquad \text{minimize}\left\{\frac{1}{2}\sum_{i\in I}(\lambda_i - \alpha_i)^2 : \sum_{i\in I}\lambda_i = 1,\ \lambda_i \geqslant 0,\ i\in I\right\}. \tag{4.37}$$

Then to determine $\bar{\lambda}_{new}$, we need to find the solutions $(\bar{\lambda}_{new}^h)_i$, $i\in Q_h$ as projections onto $\bar{M}_h$ of $\bar{\bar{\lambda}}^h = (\bar{\bar{\lambda}}_i^h,\ i\in Q_h)$ through each of the $|H|$ separable Problems $P(\bar{\bar{\lambda}}^h, Q_h)$. Thus, henceforth in this section, we will consider only one such $h\in H$. Theorem 4 below is the basis of a finitely convergent iterative scheme to solve Problem $P(\bar{\bar{\lambda}}^h, Q_h)$.

THEOREM 4: Consider the solution of Problem $P(\beta^k, I_k)$, where $\beta^k = (\beta_1^k, i \in I_k)$, with $|I_k| \geq 1$. Define

$$\rho_k = \left(1 - \sum_{i \in I_k} \beta_i^k\right) \Big/ |I_k| \tag{4.38}$$

and let

$$\bar{\beta}^k = \beta^k + (\rho_k)\, l_k \tag{4.39}$$

where l_k denotes a vector of $|I_k|$ elements, each equal to unity. Further, define

$$I_{k+1} = \{i \in I_k : \bar{\beta}_1^k > 0\}. \tag{4.40}$$

Finally, let β^{k+1} defined below be a subvector of $\bar{\beta}^k$,

$$\beta^{k+1} = (\beta_i^{k+1},\ i \in I_{k+1}) \tag{4.41}$$

where, $\beta_i^{k+1} = \bar{\beta}_i^k$, $i \in I_{k+1}$. Now suppose that $\hat{\beta}^{k+1}$ solves $P(\beta^{k+1}, I_{k+1})$.

(a) If $\bar{\beta}^k \geqslant 0$, then $\bar{\beta}^k$ solves $P(\beta^k, I_k)$.

(b) If $\bar{\beta}^k \not\geqslant 0$, then β solves $P(\beta^k, I_k)$, where β has components given by

$$\beta_i = \begin{cases} \hat{\beta}_i^{k+1}, & \text{if } i \in I_{k+1} \text{ for each } i \in I_k. \\ 0 & \text{otherwise} \end{cases} \tag{4.42}$$

PROOF: For the sake of convenience, let $RP(\alpha, I)$ denote the problem obtained by relaxing the nonnegativity restrictions in $P(\alpha, I)$. That is, let

$$RP(\alpha, I): \qquad \text{minimize} \left\{ \frac{1}{2} \sum_{i \in I} (\lambda_i - \alpha_i)^2 : \sum_{i \in I} \lambda_i = 1 \right\}.$$

First of all, note from Equations (4.38), (4.39) that $\bar{\beta}^k$ solves $RP(\beta^k, I_k)$ since $\bar{\beta}^k$ is the projection of β^k onto the linear manifold

$$\left\{ \lambda = (\lambda_i, i \in I_k) : \sum_{i \in I_k} \lambda_i = 1 \right\} \tag{4.43}$$

which is the feasible region of $RP(\beta^k, I_k)$. Thus, $\bar{\beta}^k \geqslant 0$ implies that $\bar{\beta}^k$ also solves $P(\beta^k, I_k)$. This proves part (a).

Next, suppose that $\bar{\beta}^k \not\geqslant 0$. Observe that β is feasible to $P(\beta^k, I_k)$ since from (4.42), we get $\beta \geqslant 0$ and $\sum_{i \in I_k} \beta_i = \sum_{i \in I_{k+1}} \hat{\beta}^{k+1} = 1$ as $\hat{\beta}^{k+1}$ solves $P(\beta^{k+1}, I_{k+1})$.

Now, consider any $\lambda = (\lambda_i, i \in I_k)$ feasible to $P(\beta^k, I_k)$. Then, by the Pythagorem Theorem, since $\bar{\beta}^k$ is the projection of β^k onto (4.43), we get

$$||\lambda - \beta^k||^2 = ||\lambda - \bar{\beta}^k||^2 + ||\bar{\beta}^k - \beta^k||^2.$$

Hence, the optimal solution to $P(\bar{\beta}^k, I_k)$ is also optimal to $P(\beta^k, I_k)$. Now, suppose that we can show that the optimal solution to Problem $P(\bar{\beta}^k, I_k)$ must satisfy

$$\lambda_i = 0 \text{ for } \notin I_{k+1}. \tag{4.44}$$

Then, noting (4.41), (4.42), and using the hypothesis that $\hat{\beta}^{k+1}$ solves $P(\beta^{k+1}, I_{k+1})$, we will have established part (b). Hence, let us prove that (4.44) must hold. Towards this end, consider the following Kuhn-Tucker equations for Problem $P(\bar{\beta}^k, I_k)$ with t and w_i, $i \in I_k$ as the appropriate lagrangian multipliers:

(4.46) $$\sum_{i \in I_k} \lambda_i = 1,\ \lambda_i \geqslant 0 \text{ for each } i \in I_k$$

(4.47) $$(\lambda_i - \bar{\beta}_i^k) + t - w_i = 0 \text{ and } w_i \geqslant 0 \text{ for each } i \in I_k$$

(4.48) $$\lambda_i w_i = 0 \text{ for each } i \in I_k.$$

Now, since $\sum_{i \in I_k} \bar{\beta}_i^k = 1$, we get from (4.45), (4.46) that

$$t = \sum_{i \in I_k} w_i \Big/ |I_k| \geqslant 0.$$

But from (4.46), (4.47), and (4.48) we get for each $i \in I_k$,

$$0 = w_i \lambda_i = \lambda_i (\lambda_i + t - \bar{\beta}_i^k)$$

which implies that for each $i \in I_k$, we must have,

either $\lambda_i = 0$, whence from (4.46), $w_i = t - \bar{\beta}_i^k$ must be nonnegative

or $\lambda_i = \bar{\beta}_i^k - t$, whence from (4.46), $w_i = 0$.

In either case above, noting (4.45), if $\bar{\beta}_i^k \leqslant 0$, that is, if $i \notin I_{k+1}$, we must have $\lambda_i = 0$. This completes the proof.

Using Theorem 4, one may easily validate the following procedure for finding $\bar{\lambda}_{new}^h$ of Equation (4.33), given $\bar{\bar{\lambda}}^h$. This procedure has to be repeated separately for each $h \in H$.

Initialization

Set $k = 0, \beta^0 = \bar{\bar{\lambda}}^h, I_0 = Q_h$. Go to Step 1.

Step 1

Given β^k, I_k, determine ρ_k and $\bar{\beta}^k$ from (4.38), (4.39). If $\bar{\beta}^k \geqslant 0$, then terminate with $\bar{\lambda}_{new}^h$ having components given by

$$(\bar{\lambda}_{new}^h)_i = \begin{cases} \bar{\beta}_i^k \text{ if } i \in I_k \\ 0 \text{ otherwise.} \end{cases}$$

Otherwise, proceed to Step 2.

Step 2

Define I_{k+1}, β^{k+1} as in Equations (4.40), (4.41), increment k by one and return to Step 1.

Note that this procedure is finitely convergent as it results in a strictly decreasing, finite sequence $|I_k|$ satisfying $|I_k| \geqslant 1$ for each k, since $\sum_{i \in I_k} \bar{\beta}^k = 1$ for each k.

EXAMPLE: Suppose we want to project $\bar{\bar{\lambda}}^h = (-2,3,1,2)$ on to $\Lambda \subset R^4$. Then the above procedure yields the following results.

Initialization

$k = 0,\ \beta^0 = (-2, 3, 1, 2),\ I_0 = \{1, 2, 3, 4\}$.

Step 1

$$\rho_0 = -3/4,\ \bar{\beta}^0 = \left(-\frac{11}{4},\ \frac{9}{4},\ \frac{1}{4},\ \frac{5}{4}\right)$$

Step 2

$$k = 1,\ I_1 = \{2,3,4\},\ \beta^1 = \left(\frac{9}{4},\ \frac{1}{4},\ \frac{5}{4}\right)$$

Step 1

$$\rho_1 = -\frac{11}{12},\ \bar{\beta}^1 = \left(\frac{4}{3},\ -\frac{2}{3},\ \frac{1}{3}\right)$$

Step 2

$$k = 2,\ I_2 = \{2,4\},\ \beta^2 = \left(\frac{4}{3},\ \frac{1}{3}\right)$$

Step 1

$$\rho_2 = -\frac{1}{3},\ \bar{\beta}^2 = (1,0) \geqslant 0$$

Thus, $\bar{\lambda}^h_{new} = (0,1,0,0)$.

4.1.4 A Second Sufficient Condition for Termination

As indicated earlier in Section 4.1.2, we will now derive a second sufficient condition on $\bar{w}$ for $\bar{\lambda}$ to solve PD_2. For this purpose, consider the following lemma:

LEMMA 3: Let $\bar{\lambda} \in \Lambda$ be given and suppose we obtain $\bar{w}$ using Equations (4.22) through (4.27). Let $\hat{w}$ solve the problem.

$$PR_h\text{: minimize}\left\{\frac{1}{2}\sum_{i\in Q_h}(\bar{w}_i^h - w_i^h)^2 : \sum_{i\in Q_h} w_i^h = 0,\ w_i^h \leqslant 0 \text{ for } i\in J_h\right\} \text{ for each } h\in H$$

where,

$$J_h = \{i\in Q_h : \bar{\lambda}_i^h = 0\},\ h\in H. \tag{4.49}$$

Then, if $\hat{w} = 0$, $\bar{\lambda}$ solves Problem PD_2.

PROOF. Since $\hat{w} = 0$ solves PR_h, $h\in H$, we have for each $h\in H$,

$$\sum_{i\in Q_h}(\bar{w}_i^h)^2 \leqslant \sum_{i\in Q_h}(\bar{w}_i^h - w_i^h)^2 \tag{4.50}$$

for all w_i^h, $i\in Q_h$ satisfying $\sum_{i\in Q_h} w_i^h = 0$, $w_i^h \leqslant 0$ for $i\in J_h$. Given any $\lambda \in \Lambda$ and given any $\mu > 0$ define,

$$w_i^h = (\bar{\lambda}_i^h - \lambda_i^h)/\mu,\ i\in Q_h,\ h\in H. \tag{4.51}$$

Then, $\sum_{i \in Q_h} w_i^h = 0$ for each $h \in H$ and since $\bar{\lambda}_i^h = 0$ for $i \in J_h$, $h \in H$, we get $w_i^h \leqslant 0$ for $i \in J_h$, $h \in H$. Thus, for any $\lambda \in \Lambda$, by substituting (4.51) into (4.50), we have,

$$\mu^2 \sum_{i \in Q_h} (\bar{w}_i^h)^2 \leqslant \sum_{i \in Q_h} (\lambda_i^h - \bar{\lambda}_i^h + \mu \bar{w}_i^h)^2 \quad \text{for each } h \in H. \tag{4.52}$$

But Equation (4.52) implies that for each $h \in H$, $\lambda^h = \bar{\lambda}^h$ solves the problem

$$\text{minimize} \left\{ \sum_{i \in Q_h} [\lambda_i^h - (\bar{\lambda}_i^h - \mu \bar{w}_i^h)]^2 : \sum_{i \in Q_h} \lambda_i^h = 1,\ \lambda_i^h \geqslant 0\ i \in Q_h \right\} \text{ for each } h \in H.$$

In other words, the projection $P_\Lambda(\bar{\lambda} - \bar{w}\mu)$ of $(\bar{\lambda} - \bar{w}\mu)$ onto Λ is equal to $\bar{\lambda}$ for any $\mu = 0$.

In view of Poljak's result [18,19], since $\bar{w}$ is a subgradient of $f(\cdot)$ at $\bar{\lambda}$, then $\bar{\lambda}$ solves PD_2. This completes the proof.

Note that Lemma 3 above states that if the "closest" feasible direction $-w$ to $-\bar{w}$ is a zero vector, then $\bar{\lambda}$ solves PD_2. Based on this result, we derive through Lemma 4 below a second sufficient condition for $\bar{\lambda}$ to solve PD_2.

LEMMA 4: Suppose $w = 0$ solves Problems PR_h, $h \in H$ as in Lemma 3. Then for each $h \in H$, we must have

$$(a)\ \bar{w}_i^h = t_h, \text{ a constant, for each } i \notin J_h \tag{4.53}$$

$$(b)\ \bar{w}_i^h \leqslant t_h \text{ for each } i \in J_h$$

where J_h is given by Equation (4.49).

PROOF: Let us write the Kuhn-Tucker conditions for Problem PR_h, for any $h \in H$. We obtain

$$(w_i^h - \bar{w}_i^h) + t_h = 0 \text{ for } i \notin J_h$$

$$(w_i^h - \bar{w}_i^h) + t_h - u_i^h = 0 \text{ for } i \in J_h$$

$$u_i^h \geqslant 0,\ i \in J_h,\ u_i^h w_i^h = 0\ \ i \in J_h,\ t_h \text{ unrestricted}$$

$$\sum_{i \in Q_h} w_i^h = 0,\ w_i^h \geqslant 0 \text{ for } i \in J_h.$$

If $w = 0$ solves PR_h, $h \in H$, then since PR_h has a convex objective function and linear constraints, then there must exist a solution to

$$\bar{w}_i^h = t_h \text{ for each } i \notin J_h$$

and

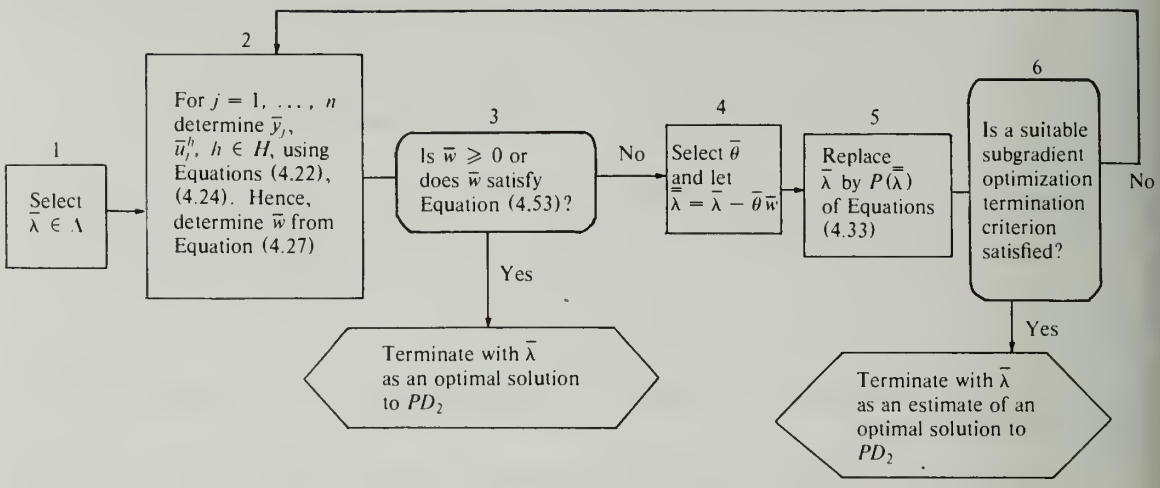
$$u_i^h = (t_h - \bar{w}_i^h) \geqslant 0 \text{ for each } i \in J_h.$$

This completes the proof.

Thus Equation (4.53) gives us another sufficient condition for $\bar{\lambda}$ to solve PD_2. We illustrate the use of this condition through an example in Section 4.1.7.

4.1.5 Schema of an Algorithm to Solve Problem PD_2

The procedure is depicted schematically below. In block 1, an arbitrary or preferably, a good heuristic solution $\bar{\lambda} \in \Lambda$ is sought. For example, one may use $\bar{\lambda}_i^h = 1/|Q_h|$ for each $i \in Q_h$, for $h \in H$. For blocks 4 and 6, we recommend the procedural steps proposed by Held, Wolfe and Crowder [12] for the subgradient optimization scheme.


1
Select $\bar{\lambda} \in \Lambda$
2
For $j = 1, \ldots, n$ determine $\bar{y}_j$, $\bar{u}_j^h$, $h \in H$, using Equations (4.22), (4.24). Hence, determine $\bar{w}$ from Equation (4.27)
3
Is $\bar{w} \geq 0$ or does $\bar{w}$ satisfy Equation (4.53)?
No
4
Select $\bar{\theta}$ and let $\bar{\bar{\lambda}} = \bar{\lambda} - \bar{\theta}\bar{w}$
5
Replace $\bar{\lambda}$ by $P(\bar{\bar{\lambda}})$ of Equations (4.33)
6
Is a suitable subgradient optimization termination criterion satisfied?
No
Yes
Terminate with $\bar{\lambda}$ as an optimal solution to PD_2
Yes
Terminate with $\bar{\lambda}$ as an estimate of an optimal solution to PD_2

4.1.6 Derivation of a Good Subgradient Direction

In our discussion in Section 4.1.1, we saw that given a $\lambda \in \Lambda$ of Equation (4.28), we were able to uniquely determine $\bar{y}_j$, $j = 1, \ldots, n$ through Equation (4.22). Thereafter, once we fixed values $\bar{u}_j^h$ for u_j^h, $j = 1, \ldots, n$, $h \in H$ satisfying Equation (4.24), we were able to uniquely determine values for the other variables in the Kuhn-Tucker System using Equations (4.26), (4.27). Moreover, the only choice in determining $\bar{u}_j^h$, $j = 1, \ldots, n$, $h \in H$ arose in case $|H_j| \geq 2$ for some $j \in \{1, \ldots, n\}$ in Equation (4.25). We also established that no matter what feasible values we selected for u_j^h, $j \in \{1, \ldots, n\}$, $h \in H$, the corresponding vector w obtained was a subgradient direction. In order to select the best such subgradient direction, we are interested in finding a vector $\bar{w}$ which has the smallest euclidean norm among all possible vectors corresponding to the given solution $\bar{\lambda} \in \Lambda$. However, this problem is not easy to solve. Moreover, since this step will merely be a subroutine at each iteration of the proposed scheme to solve PD_2, we will present a heuristic approach to this problem.

Towards this end, let us define for convenience, mutually exclusive but not uniquely determined sets $N_h, h \in H$ as follows:

(4.54) $\quad N_h \subset \{j \in \{1, \ldots, n\}: h \in H_j \text{ of Equation (4.23)}\}$

(4.55) $\quad N_i \cap N_j = \{\phi\}$ for any i, $j \in H$ and $\bigcup_{h \in H} N_h = \{j \in \{1, \ldots, n\}: \bar{y}_j > 0\}$.

In other words, we take each $j \in \{1, \ldots, n\}$ which has $\bar{y}_j > 0$, and assign it to some $h \in H_j$, that is, assign it to a set N_h, where $h \in H_j$. Having done this, we let

(4.56) $$\bar{u}_j^h = \begin{cases} 2\bar{y}_j \text{ if } j \in N_h \\ 0 \text{ otherwise} \end{cases} \quad \text{for each } j \in \{1, \ldots, n\},\ h \in H.$$

Note that Equation (4.56) yields values $\bar{u}_j^h$ for u_j^h, $j \in \{1, \ldots, n\}$, $h \in H$ which are feasible to (4.24). Hence, having defined sets N_h, $h \in H$ as in Equations (4.54), (4.55), we determine $\bar{u}_j^h$, $j \in \{1, \ldots, n\}$, $h \in H$ through (4.56) and hence $\bar{w}$ through (4.27).

Thus, the proposed heuristic scheme commences with a vector w obtained through an arbitrary selection of sets N_h, $h \in H$ satisfying Equations (4.54), (4.55). Thereafter, we attempt to improve (decrease) the value of $w^t w$ in the following manner. We consider in turn each $j \in \{1, \ldots, n\}$ which satisfies $|H_j| \geq 2$ and move it from its current set N_{h_j}, say, to another set

N_h with $h \in H_j$, $h \neq h_j$, if this results in a decrease $w^t w$. If no such single movements result in a decrease in $w^t w$, we terminate with the incumbent solution w as the sought subgradient direction. This procedure is illustrated in the example given below.

4.1.7 Illustrative Example

The intention of this subsection is to illustrate the scheme of the foregoing section for determining a good subgradient direction as well as the termination criterion of Section 4.1.4.

Thus, let $H = \{1,2\}$, $n = 3$, $|Q_1| = |Q_2| = 3$ and consider the constraint sets

$$S_1 = \left\{ \begin{array}{rl} x: & 2x_1 - 3x_2 + x_3 \geqslant 1 \\ & -x_1 + 2x_2 + 3x_3 \geqslant 1 \\ & 3x_1 - x_2 - x_3 \geqslant 1 \\ & x_1, x_2, x_3 \geqslant 0 \end{array} \right\} \text{ and } S_2 = \left\{ \begin{array}{rl} x: & 3x_1 - x_2 - x_3 \geqslant 1 \\ & 2x_1 + x_2 - 2x_3 \geqslant 1 \\ & -x_1 + 3x_2 + 3x_3 \geqslant 1 \\ & x_1, x_2, x_3 \geqslant 0 \end{array} \right\}$$

Further, suppose we are currently located at a point $\bar{\lambda}$ with

$$\bar{\lambda}_1^1 = 0,\ \bar{\lambda}_2^1 = 5/12,\ \bar{\lambda}_3^1 = 7/12;\ \bar{\lambda}_1^2 = 7/12,\ \bar{\lambda}_2^2 = 0,\ \bar{\lambda}_3^2 = 5/12.$$

Then the associated surrogate constraints are

$$\frac{4}{3}x_1 + \frac{1}{4}x_2 + \frac{2}{3}x_3 \geqslant 1 \text{ for } h = 1$$

(4.57)

$$\frac{4}{3}x_1 + \frac{2}{3}x_2 + \frac{2}{3}x_3 \geqslant 1 \text{ for } h = 2.$$

Using Equations (4.22), (4.25), we find

$$\bar{y}_1 = \frac{4}{3} \text{ with } H_1 = \{1,2\},\ \bar{y}_2 = \frac{2}{3} \text{ with } H_2 = \{2\} \text{ and } \bar{y}_3 = \frac{2}{3} \text{ with } H_3 = \{1,2\}.$$

Note that the possible combinations of N_1 and N_2 are as follows:

(i) $N_1 = \{1\}$, $N_2 = \{2,3\}$,

(ii) $N_1 = \{\phi\}$, $N_2 = \{1,2,3\}$,

(iii) $N_1 = \{1,3\}$, $N_2 = \{2\}$, and

(iv) $N_1 = \{3\}$, $N_2 = \{1,2\}$.

A total enumeration of the values of u obtained for these sets through (4.56) and the corresponding values for w are shown below.

		$u_j^h,\ j \in \{1,\ldots,n\}$						$w_i^h,\ i \in Q_h,\ h \in H$						
N_1	N_2	u_1^1	u_2^1	u_3^1	u_1^2	u_2^2	u_3^2	w_1^1	w_2^1	w_3^1	w_1^2	w_2^2	w_3^2	$w^t w$
{1}	{2,3}	8/3	0	0	0	4/3	4/3	16/9	−56/9	40/9	−40/9	−28/9	56/9	129.78
{ϕ}	{1,2,3}	0	0	0	8/3	4/3	4/3	0	0	0	0	−4/3	0	1.78
{1,3}	{2}	8/3	0	4/3	0	4/3	0	20/9	−28/9	20/9	−20/9	4/9	28/9	34.37
{3}	{1,2}	0	0	4/3	8/3	4/3	0	−4/9	28/9	−20/9	20/9	20/9	−28/9	34.37

Thus, according to the proposed scheme, if we commence with $N_1 = \{1\}$, $N_2 = \{2,3\}$, then picking $j = 1$ which has $|H_j| = 2$, we can move $j = 1$ into N_2 since $2 \in H_1$. This leads to an improvement. As one can see from above, no further improvement is possible. In fact, the

best solution shown above is accessible by the proposed scheme by all except the third case which is a "local optimal".

We now illustrate the sufficient termination condition of Section 4.1.4. The vector $\bar{w}$ obtained above is $(\overset{h=1}{0,0,0}|0,-4/3,0)$. Further the vector $\bar{\lambda}$ is $(\overset{h=1}{0,5/12,7/12}|\overset{h=2}{7/12,0,5/12})$. Thus, even though $\bar{w} \geq 0$, we see that the conditions (4.53) of Lemma 6 are satisfied for each $h \in H = \{1,2\}$ and thus the given $\bar{\lambda}$ solves PD_2.

The disjunctive cut (3.8) derived with this optimal solution $\bar{\lambda}$ is obtained through (4.57) as

$$\frac{4}{3}x_1 + \frac{2}{3}x_2 + \frac{2}{3}x_3 \geq 1. \tag{4.58}$$

It is interesting to compare this cut with that obtained through the parameter values $\bar{\lambda}_i^h = 1/|Q_h|$ for each $i \in Q_h$ as recommended by Balas [1,2]. This latter cut is

$$\frac{4}{3}x_1 + x_2 + x_3 \geq 1. \tag{4.59}$$

Observe that (4.58) uniformly dominates (4.59).

4.2 Maximizing the Rectilinear Distance Between the Origin and the Disjunctive Cut

In this section, we will briefly consider the case where one desires to use rectilinear instead of euclidean distances. Extending the developments of Sections 2, 3 and 4.1, one may easily see that the relevant problem is

$$\text{minimize } \{\underset{j \in \{1,\ldots,n\}}{\text{maximum}}\ y_j\text{: constraints (4.12), (4.13), (4.14) are satisfied}\}.$$

The reason why we consider this formulation is its intuitive appeal. To see this, note that the above problem is separable in $h \in H$ and may be rewritten as

$$PD_1\text{: minimize} \left\{ \xi^h\text{: } \xi^h \geq \sum_{i \in Q_h} \lambda_i^h a_{ij}^h \text{ for each } j = 1, \ldots, n, \sum_{i \in Q_h} \lambda_i^h = 1, \lambda_i^h \geq 0 \right.$$
$$\left. \text{for } i \in Q_h,\ \xi^h \geq 0 \right\} \text{ for each } h \in H.$$

Thus, for each $h \in H$, PD_1 seeks λ_i^h, $i \in Q_h$ such that the largest of the surrogate constraint coefficients is minimized. Once such surrogate constraints are obtained, the disjunctive cut (3.8) is derived using the principles of Section 3.

As far as the solution of Problem PD_1 is concerned, we merely remark that one may either solve it as a linear program or rewrite it as the minimization of a piecewise linear convex function subject to linear constraints and use a subgradient optimization technique.

BIBLIOGRAPHY

[1] Balas, E., "Intersection Cuts from Disjunctive Constraints," Management Science Research Report, No. 330, Carnegie-Mellon University (1974).

[2] Balas, E., "Disjunctive Programming: Cutting Planes from Logical Conditions in Nonlinear Programming," O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, editors, Academic Press, New York (1975).

[3] Balas, E., "Disjunctive Programming: Facets of the Convex Hull of Feasible Points," Management Science Research Report, No. 348, Carnegie-Mellon University, (1974).

[4] Bazaraa, M.S. and C.M. Shetty, "Nonlinear Programming: Theory and Algorithms," John Wiley and Sons, New York (1979).

[5] Burdet, C., "Elements of a Theory in Non-Convex Programming," Naval Research Logistics Quarterly, *24*, 47-66 (1977).

[6] Burdet, C. "Convex and Polaroid Extensions," Naval Research Logistics Quarterly, *24*, 67-82 (1977).

[7] Dem'janov, V.F., "Seeking a Minimax on a Bounded Set," Soviet Mathematics Doklady, *11*, 517-521 (1970) (English Translation).

[8] Glover, F., "Convexity Cuts for Multiple Choice Problems," Discrete Mathematics, *6*, 221-234 (1973).

[9] Glover, F., "Polyhedral Convexity Cuts and Negative Edge Extensions," Zeitschrift für Operations Research, *18*, 181-186 (1974).

[10] Glover, F., "Polyhedral Annexation in Mixed Integer and Combinatorial Programming," Mathematical Programming, *8*, 161-188 (1975). See also MSRS Report 73-9, University of Colorado (1973).

[11] Glover, F., D. Klingman and J. Stutz, "The Disjunctive Facet Problem: Formulation and Solutions Techniques," Management Science Research Report, No. 72-10, University of Colorado (1972).

[12] Held, M., P. Wolfe and H.D. Crowder, "Validation of Subgradient Optimization," Mathematical Programming, *6*, 62-88 (1974).

[13] Jeroslow, R.G., "The Principles of Cutting Plane Theory: Part I," (with an addendum), Graduate School of Industrial Administration, Carnegie-Mellon University (1974).

[14] Jeroslow, R.G., "Cutting Plane Theory: Disjunctive Methods," Annals of Discrete Mathematics, *1*, 293-330 (1977).

[15] Karlin, S., "Mathematical Methods and Theory in Games, Programming and Economics," *1*, Addison-Wesley Publishing Company, Reading, Mass. (1959).

[16] Majthay, A. and A. Whinston, "Quasi-Concave Minimization Subject to Linear Constraints," Discrete Mathematics, *9*, 35-59 (1974).

[17] Owen, G., "Cutting Planes for Programs with Disjunctive Constraints," Optimization Theory and Its Applications, *11*, 49-55 (1973).

[18] Poljak, B.T., "A General Method of Solving Extremum Problems," Soviet Mathematics Doklady, *8*, 593-597 (1967). (English Translation).

[19] Poljak, B.T., "Minimization of Unsmooth Functionals," USSR Computational Mathematics and Mathematical Physics, *9*, 14-29 (1969). (English Translation).

[20] Vaish, H. and C. M. Shetty, "A Cutting Plane Algorithm for the Bilinear Programming Problem," Naval Research Logistics Quarterly, *24*, 83-94 (1975).

THE ROLE OF INTERNAL STORAGE CAPACITY IN FIXED CYCLE PRODUCTION SYSTEMS

B. Lev*

Temple University
Philadelphia, Pennsylvania

D. I. Toof

Ernst & Ernst
Washington, D.C.

ABSTRACT

The reliability of a serial production line is optimized with respect to the location of a single buffer. The problem was earlier defined and solved by Soyster and Toof for the special case of an even number of machines all having equal probability of failure. In this paper we generalize the results for any number of machines and remove the restriction of identical machine reliabilities. In addition, an analysis of multibuffer systems is presented with a closed form solution for the reliability when both the number of buffers and their capacity is limited. For the general multibuffer system we present an approach for determining system reliability.

1. INTRODUCTION

Several types of production line models appear in the literature. Each one is a realization of a different real life situation. A summary of the various types and the differences in the mechanism of product flow among them appears in Buzacott [5], Koenigsberg [9], Toof [14] or Buxey et al [1]. Recently Soyster and Toof [13] defined a serial production line, which is the model analyzed in this paper.

The mechanism of product flow in a serial production line is described via Figure 1. An unlimited source of raw material exists before machine 1. If machine 1 is capable of working (i.e., not failed), an operator takes a unit of raw material and processes it on machine 1, after which he moves to machine 2 and processes it on machine 2, if machine 2 is capable of working. He proceeds analagously until machine N where a finish product is completed. Let T_i be the process time on machine i. Then the cycle time of the system $T = \sum_{i=1}^{N} T_i$. Let q_i be the probability that at any cycle T machine i is capable of working and $p_i = 1 - q_i$ the probability of failing. The serial production line with no buffer must stop working if any of the individual machines on the line fails. The placement of a single buffer of capacity M after machine i alleviates this situation. If any of the first i machines fail and the buffer is not empty, machines

*This study was done when the author was at the Department of Energy, Washington, D.C. under the provisions of the Intergovernmental Personnel Act.

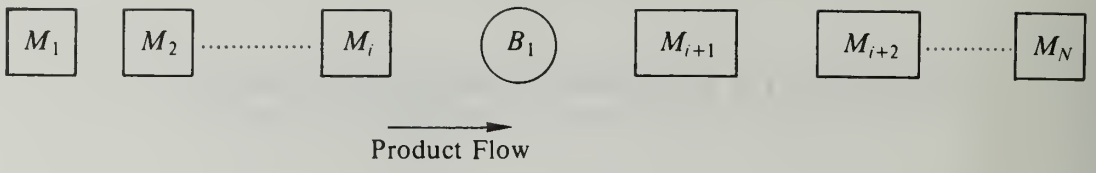


FIGURE 1. Serial production line with N machines and a single buffer

$i + 1, i + 2, \ldots, N$ can still function. Conversely, if any of the machines $i + 1, \ldots, N$ fail and the buffer is not full, the first i machines may still work and produce a semifinished good to be stored in the buffer. One obviously would like to identify the optimal placement of this buffer. Soyster and Toof [13] proved that if there are an even number of machines, all identically reliable ($q_i = q \ \forall \ i$) then the optimal placement of the buffer is exactly in the middle of the line. In section 2 we generalize these results for any number of machines not necessarily identically reliable. Specifically, we prove that the optimal placement of a single buffer is at a place which minimizes the absolute value of the difference between the reliability of the two parts of the line separated by the buffer.

The optimal location i^* is determined from (1)

$$(1) \qquad \prod_{j=1}^{i^*} q_j - \prod_{j=i^*+1}^{N} q_j = \operatorname*{Min}_{1 \leqslant i \leqslant N} \left| \prod_{j=1}^{i} q_j - \prod_{j=i+1}^{N} q_j \right|.$$

A more difficult question is the optimal locations of several buffers. In section 3 we analyze a special case of a two buffer system, each buffer having a capacity of one unit. In section 4 we present an approach that can be used for any number of buffers with any capacity. The approach we suggest is efficient as long as the number of buffers and their capacity remains relatively small.

2. OPTIMAL LOCATION OF A SINGLE BUFFER

Let a single buffer with capacity M be placed after machine i. Let $\alpha_i = \prod_{j=1}^{i} q_j$, $\beta_i = \prod_{j=i+1}^{N} q_j$, $\rho_i = (\alpha_i - \alpha_i\beta_i)/(\beta_i - \alpha_i\beta_i)$, and let X_n be the number of units in the buffer at the beginning of cycle n. Soyster and Toof [13] have shown that X_n defines a finite Markov Chain, presented its transition matrix and found that the reliability $R(i)$ of the line is given by (2) and (3):

$$(2) \qquad R(i) = \beta_i\alpha_i + \beta_i(1 - \alpha_i)\frac{\rho_i - \rho_i^{M+1}}{1 - \rho_i^{M+1}} \qquad \text{if } \alpha_i \neq \beta_i$$

$$(3) \qquad R(i) = \beta_i\alpha_i + \beta_i(1 - \alpha_i)\frac{M}{M+1} \qquad \text{if } \alpha_i = \beta_i.$$

One has to maximize $R(i)$ with respect to i, that is, to identify the optimal location of the buffer within the line. Since $\alpha_i\beta_i = \prod_{j=1}^{i} q_j \prod_{j=i+1}^{N} q_j = \prod_{j=1}^{N} q_j$ is a constant and does not affect the location of the buffer, one can simply ignore this term from (2) and (3) in the optimization phase. Thus, we want to find i^* that maximizes $R(i)$ or:

$$(4) \qquad R(i^*) = \operatorname*{Max}_i R(i) = \operatorname*{Max}_i \begin{cases} \beta_i(1 - \alpha_i)\dfrac{\rho_i - \rho_i^{M+1}}{1 - \rho_i^{M+1}} & \text{if } \alpha_i \neq \beta_i \\ \beta_i(1 - \alpha_i)\dfrac{M}{M+1} & \text{if } \alpha_i = \beta_i. \end{cases}$$

The approach we take to solve (4) for i^* is to show that $R(i)$ is strictly increasing with α_i for $\alpha_i < \beta_i$ and strictly decreasing with α_i for $\alpha_i > \beta_i$; that $\alpha_i = \beta_i$ occurs when $R(i)$ reaches its maximum value; and that $R(i)$ is symmetric about the point i^* where $\alpha_{i^*} = \beta_{i^*}$.

Let

$$R(i) = (\beta_i - \alpha_i\beta_i)\frac{\rho_i^{M+1} - \rho_i}{\rho_i^{M+1} - 1} \qquad \alpha_i \neq \beta_i \tag{5}$$

when $\alpha_i = \beta_i$, $\rho_i = 1$ and (5) becomes (6)

$$R(i) = (\beta_i - \alpha_i\beta_i)\frac{M}{M+1} \qquad \alpha_i = \beta_i. \tag{6}$$

Note in (6) as M becomes large the total reliability of the line, which is equal to $\alpha_i\beta_i + R(i)$, approaches β_i. That is, the two segments of the line become independent of each other.

In this section the general strategy is to show that if $\alpha_i > \beta_i$ or $\alpha_i < \beta_i$ then the reliability of (5) is smaller than the reliability of (6). Hence, we treat α_i as a continuous variable and show that the derivative of (5) with respect to α_i is positive for $\alpha_i < \beta_i$ and negative for $\alpha_i > \beta_i$.

The derivative of $R(i)$ with respect to α_i is:

$$\frac{dR(i)}{d\alpha_i} = \frac{1}{\rho_i^{M+1}-1}\left\{\left[\frac{-\alpha_i\beta_i}{\alpha_i^2}\right](\rho_i^{M+1}-\rho_i) + \left[\frac{(M\rho_i^{M+1}-(M+1)\rho_i^{M}+1)}{(\rho_i^{M+1}-1)^2}\right]\cdot\left[1+\frac{\rho_i\beta_i}{\alpha_i}\right]\right\}.$$

LEMMA 1: The additional reliability function $R(i)$, is strictly increasing with respect to α_i over the range $\left[0, \left(\prod_{i=1}^{N} q_i\right)^{1/2}\right]$, and strictly decreasing with respect to α_i over the range $\left(\left(\prod_{i=1}^{N} q_i\right)^{1/2}, 1\right]$. That is, if $0 \leqslant \alpha_i < \beta_i$, then $\frac{dR(i)}{d\alpha_i} > 0$. Conversely, if $\beta_i < \alpha_i \leqslant 1$, then $\frac{dR(i)}{d\alpha_i} < 0$. The proof can be found in [14]. (The first range is closed from the left and open from the right; the second range is open from the left and closed from the right).

THEOREM 1: The optimal placement, i^*, of a single buffer of integer capacity M in an N machine line is where $\alpha_i^* = \beta_i^*$.

PROOF: The proof of this theorem is essentially complete. We must only show that (5) is continuous at the point where $\alpha_i^* = \beta_i^*$. By definition the additional reliability attributable to the introduction of the buffer when $\alpha_i^* = \beta_i^*$ is:

$$(\beta_i^* - \alpha_i^*\beta_i^*)\frac{M}{M+1}.$$

As $\alpha_i \to \beta_i$ $\rho_i \to 1$ so that in (5) the limit of the steady state probability as $\alpha_i \to \beta_i$ is of the indeterminate form 0/0. However, an application of L'Hospital's rule shows that:

$$\lim_{\alpha_i \to \beta_i} \frac{\rho_i^{M+1}-\rho_i}{\rho_i^{M+1}-1} = \frac{M}{M+1}$$

and thus the continuity is proven.

Theorem 1 defines an optimal though not necessarily feasible solution to the problem of buffer placement. The condition $\alpha_i = \beta_i$ may be impossible to satisfy. In the remainder of this section we examine the symmetry of the reliability function defined by Equation (5), develop a simple criterion that provides the best feasible solution and, lastly, we examine the special case of identical machine reliability, i.e., $q_i = q \ \forall \ i$.

LEMMA 2: Given K_1 and K_2 continuous variables such that $\alpha_{K_1} - \beta_{K_1} = \beta_{K_2} - \alpha_{K_2}$. Then $\rho_{K_1} \cdot \rho_{K_2} = 1$.

PROOF: Recall that $\alpha_i \beta_i = \prod_{i=1}^{N} q_i = Q$ a constant for all i. Thus the condition $\alpha_{K_1} - \beta_{K_1} = \beta_{K_2} - \alpha_{K_2}$ may be rewritten $\alpha_{K_1} - \frac{Q}{\alpha_{K_1}} = \frac{Q}{\alpha_{K_2}} - \alpha_{K_2}$. This implies that:

$$\alpha_{K_1} + \alpha_{K_2} = \frac{Q(\alpha_{K_1} + \alpha_{K_2})}{\alpha_{K_1}\alpha_{K_2}} \text{ or that } \alpha_{K_1}\alpha_{K_2} = Q.$$

Similarly, one obtains the result that $\beta_{K_1}\beta_{K_2} = Q$. We want to show that $\rho_{K_1} \cdot \rho_{K_2} = 1$. Substituting for ρ_{K_1} and ρ_{K_2} in the definition of ρ yields:

$$\rho_{K_1}\rho_{K_2} = \frac{(\alpha_{K_1} - Q)(\alpha_{K_2} - Q)}{(\beta_{K_1} - Q)(\beta_{K_2} - Q)}.$$

We then must show that:

$$(\alpha_{K_1} - Q)(\alpha_{K_2} - Q) = (\beta_{K_1} - Q)(\beta_{K_2} - Q)$$

or that:

$$\alpha_{K_1}\alpha_{K_2} - Q(\alpha_{K_1} + \alpha_{K_2}) = \beta_{K_1}\beta_{K_2} - Q(\beta_{K_1} + \beta_2).$$

The condition $\alpha_{K_1} - \beta_{K_1} = \beta_{K_2} - \alpha_{K_2}$ infers both that $\alpha_{K_1} + \alpha_{K_2} = \beta_{K_1} + \beta_{K_2}$ and that $\alpha_{K_1}\alpha_{K_2} = \beta_{K_1}\beta_{K_2} = Q$, and thus the proof is complete.

This leads directly to the following theorem:

THEOREM 2: For a continuous argument (i), $R(i)$ is symmetric about the point i^* where $\alpha_{i^*} = \beta_{i^*}$.

The proof is in [14].

The placement of the buffer has been treated as a continuous variable. While this has led to satisfying mathematical results, in reality one must develop an optimizing criterion which is physically feasible. Unfortunately, the condition $\alpha_{i^*} = \beta_{i^*}$ does not satisfy the feasibility requirements. Rarely will i^* be integer and what, for example, is the physical interpretation of $i^* = 7.63$. To this end, it will be shown in this section that the steady state reliability of the line is maximized by placing the buffer after machine i^* (i^* integer) where i^* satisfies the following condition:

$$|\alpha_{i^*} - \beta_{i^*}| = \min_{1 \leqslant i \leqslant N} |\alpha_i - \beta_i|.$$

Note that if an integer i^* exists such that $\alpha_{i^*} = \prod_{j=1}^{i^*} q_j = \prod_{j=1^*+1}^{N} q_j = \beta_{i^*}$, it would satisfy the above criterion and be consistent with Theorem 1.

To this end observe that $|\alpha_i - \beta_i|$ is a convex function of α_i that obtains its minimum point at $\alpha_i = \beta_i = \sqrt{\alpha_i \beta_i} = \sqrt{Q}$. Thus, for

$$\alpha_i < \alpha_j < \sqrt{Q},\ |\alpha_j - \beta_j| < |\alpha_i - \beta_i|, \text{ and for } \sqrt{Q} < \alpha_j < \alpha_i,\ |\alpha_j - \beta_j| < |\alpha_i - \beta_i|.$$

THEOREM 3 (Fundamental): The optimal integer placement of a single buffer of capacity M in an N machine line is where $|\alpha_i - \beta_i|$ is minimized.

PROOF: From Theorem 1 we know that by treating i as a continuous variable the optimal placement i^* satisfies $\alpha_{i^*} = \beta_{i^*}$. If i^* is integer the theorem is evident. Assume that i^* is not integer. Examine the points $[i^*]$ and $[i^* + 1]$. From lemma 1 and the convexity of $|\alpha_i - \beta_i|$ we know that $R([i^*]) > R(K_1)$ where $\alpha_{[i^*]} > \alpha_{K_1}$ and $R([i^* + 1]) > R(K_2)$ where $\alpha_{[i^*+1]} > \alpha_{K_2}$. Thus, the only two candidate placements are $[i^*]$ and $[i^* + 1]$.

If $|\alpha_{[i^*]} - \beta_{[i^*]}| = |\alpha_{[i^*+1]} - \beta_{[i^*+1]}|$ then the theorem holds and either placement is optimal. Therefore, assume that $|\alpha_{[i^*]} - \beta_{[i^*]}| < |\alpha_{[i^*+1]} - \beta_{[i^*+1]}|$. We want to show that $R([i^*]) > R([i^* + 1])$. Assume the contrary, i.e., that $R([i^* + 1]) > R([i^*])$. From Theorem 2 we know that there exists a point K^* such that $R(K^*) = R([i^* + 1])$ and that $|\alpha_{K^*} - \beta_{K^*}| = |\alpha_{[i^*+1]} - \beta_{[i^*+1]}|$. This implies that $R(K^*) > R([i^*])$. We know that $|\alpha_{K^*} - \beta_{K^*}| > |\alpha_{[i^*]} - \beta_{[i^*]}|$ and since both α_{K^*} and $\alpha_{[i^*]}$ must be greater than $\sqrt{\alpha_i \beta_i}$ this implies that $\alpha_{K^*} > \alpha_{[i^*]}$. By Theorem 2 this would infer that $R([i^*]) > R(K^*)$ which is a contradiction. Similar results may be obtained by assuming that $|\alpha_{[i^*]} - \beta_{[i^*]}| > |\alpha_{[i^*+1]} - \beta_{[i^*+1]}|$.

Theorem 3 details a simple, yet elegant criterion for the optimal placement of a single buffer regardless of capacity so as to maximize the reliability of the system.

A Special Case: $q_i = q\ \forall\ i$.

Consider the case where $q_i = \forall\ i$. In this case:

$$\alpha_i = q^i$$

$$\beta_i = q^{N-i}.$$

It follows from Theorems 1 and 3 that if N is even, the optimal placement would be where $\alpha_i = \beta_i$ which in this case is where $q^i = q^{N-i}$ which is satisfied at $i = N/2$. This is consistent with the results developed by Soyster and Toof [13].

Assume that N is odd. Then N is of the form $2K + 1$ where K is integer and by Theorem 3 the optimal placement is either after machine K or machine $K + 1$ since:

$$|\alpha_K - \beta_K| = |q^K - q^{2K+1-K}| = |q^K - q^{K+1}|$$

$$|\alpha_{K+1} - \beta_{K+1}| = |q^{K+1} - q^{2K+1-K-1}| = |q^K - q^{K+1}|.$$

We have just completed the proof for the optimal location of a single buffer on an N machine serial line. The optimal location is for any N (even or odd) and for any q_i (both when machine reliability are identical or not identical for all machines). In the next section we generalize the model to include more than one buffer.

3. TWO BUFFERS OF CAPACITY ONE UNIT

Consider a simpler case of the general model where $N = 3k$ and $q_i = q$ for all i. The placement of two buffers separates the line into three segments. Since $N = 3k$, one may arbitrarily place the first buffer immediately after machine k and the second immediately after

machine $2k$. The placement of these two buffers has just defined the three stages of the system. Each stage may be comprised of more than one machine; for a line of $N = 3k$, each stage is comprised of k machines. The reliability of each stage is $Q_1 = Q_2 = Q_3 = q_i^k = Q$ and $P = 1 - Q$.

The two buffer system operates analogously to the one buffer system described in section 2. If all machines are up, then a unit of raw material is processed by stages one, two and three and a finished good is produced. If, for example, stage three is down, stages one and two are up and buffer two is not full, then both stages one and two operate and a semicompleted good would be stored in buffer two. If buffer two had been full and buffer one had not, then machine two would not operate; it would be blocked by the second buffer which is full. In this case only machine one would operate and a semiprocessed good would be stored in buffer one.

Define an ordered pair (X, Y) where X represents the quantity of semifinished goods in buffer one at the start of cycle t, and Y the quantity in buffer two at the start of cycle t. If we assume that the maximum capacity of both buffers one and two is one, then the pair (X, Y) may take on the following four values: (0,0), (1,0), (0,1), and (1,1). The one cycle transition probability from state $(X, Y) = (0, 0)$ to all states is:

- Both are empty at the start of cycle $t + 1$ if either all stages are up, or if stage one is down. Thus: $P[(X_{t+1}, Y_{t+1}) = (0,0) \mid (X_t, Y_t) = (0,0)] = Q^3 + P$.
- If stage one is up during cycle t but stage two is down, then a unit of raw material is processed on stage one and the semicompleted good stored in buffer one. Thus: $P[(X_{t+1}, Y_{t+1}) = (1,0) \mid (X_t, Y_t) = (0,0)] = QP$.
- If both stages one and two are up but stage three is down, then a unit of raw material is processed on both stages one and two and the semicompleted good stored in buffer two. Thus: $P[(X_{t+1}, Y_{t+1}) = (0,1) \mid (X_t, Y_t) = (0,0)] = Q^2P$.
- Lastly, note that it is impossible for (X_{t+1}, Y_{t+1}) to equal (1,1) given that $(X_t, Y_t) = (0, 0)$, as at most, one unit may be added to storage during any cycle. Thus: $P[(X_{t+1}, Y_{t+1}) = (1,1) \mid (X_t, Y_t) = (0,0)] = 0$.

One may compute the transition probabilities for all of the four possible states in an analogous manner. The complete transition matrix is presented in Figure 2.

State in t \ State in t+1	(0,0)	(1,0)	(0,1)	(1,1)
(0,0)	Q^3+P	QP	Q^2P	0
(1,0)	Q^2P	Q^3+P	QP^2	Q^2P
(0,1)	QP	Q^2P	Q^3+P^2	QP
(1,1)	0	QP	Q^2P	Q^3+P

FIGURE 2. Transition matrix — two buffer system

Let $\pi_1, \pi_2, \pi_3, \pi_4$ be the steady state probabilities of buffer states (0,0), (1,0), (0,1) and (1,1) respectively. The system is in state (0,0) with probability π_1, then a good is produced if and only if all three stages are up. This event has a probability of $Q^3\pi_1$. Similarly, with probability π_2 the system is in state (1,0), then only stages two and three must be up for a finished good to be produced. This event has probability $Q^2\pi_2$. Lastly in both state (0,1) and

(1,1), buffer two is not empty and thus the only condition for a successful cycle is that stage three must be up. These events have probability $Q\pi_3$ and $Q\pi_4$, respectively. The steady state reliability, R, of the two buffer system where the capacity of both buffer one and buffer two is one unit is equal to:

$$R = Q^3\pi_1 + Q^2\pi_2 + Q\pi_3 + Q\pi_4. \tag{7}$$

Thus, upon determining the steady state probabilities, π_1, π_2, π_3 and π_4, one has an exact formulation of the reliability of the three stage, two buffer system, where each buffer has a capacity of one unit.

From the transition matrix presented as Figure 2 and basic finite Markov Chain theory, one can calculate π_1, π_2, π_3, and π_4 in the following manner.

First, we know that in the steady state $\pi B = \pi$ where B is the one step transition matrix of the system (Figure 2) and

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4).$$

This identity yields a system of four simultaneous equations of the form

$$\pi(B - I) = 0 \tag{8}$$

where B is the form:

$$B = \begin{pmatrix} Q^3 + P & QP & Q^2P & 0 \\ Q^2P & Q^3 + P & QP^2 & Q^2P \\ QP & Q^2P & Q^3 + P^2 & QP \\ 0 & QP & Q^2P & Q^3 + P \end{pmatrix}.$$

However, (B-I) has no inverse as the rows are linearly dependent. The classical method of solution to this problem is to drop one of the identity equations of π and substitute the fact that the sum of the steady state probabilities must equal one. That is, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. Making this substitution for column 3 of B-I yields the following system of simultaneous equations: $\pi A = (0,0,1,0)$, where:

$$A = \begin{pmatrix} Q^3 + P - 1 & QP & 1 & 0 \\ Q^2P & Q^3 + P - 1 & 1 & Q^2P \\ QP & Q^2P & 1 & QP \\ 0 & QP & 1 & Q^3 + P - 1 \end{pmatrix}.$$

Thus, $\pi = (0,0,1,0)A^{-1}$ which reduces to $\pi = A_3^{-1}$ where A_3^{-1} is the third column of the inverse matrix A^{-1}. The solution to the last system of four equations and four variables is:

$$\left.\begin{aligned} \pi_1 &= (Q^2 + Q + 1)/(4Q^2 + 3Q + 5) \\ \pi_2 &= (Q^2 + Q + 2)/(4Q^2 + 3Q + 5) \\ \pi_3 &= (Q^2 + 1)/(4Q^2 + 3Q + 5) \\ \pi_4 &= (Q^2 + Q + 1)/(4Q^2 + 3Q + 5) \end{aligned}\right\}. \tag{9}$$

We are now able to directly compute the steady state reliability of a two buffer series system where each stage has identical reliability, Q, distributed Bernoulli and each buffer a capacity of one unint. We have just proved Theorem 4 which results from (7) and (9).

THEOREM 4: For the series production system described above the steady state reliability of the system R, is equal to:

$$R = \frac{Q^5 + 2Q^4 + 4Q^3 + 3Q^2 + 2Q}{4Q^2 + 3Q + 5}.$$

4. EXTENSION OF THE GENERAL MUTLIBUFFER CASE

The previous sections have laid the groundwork for our analysis of a general multistage, multibuffer system such as the one depicted in Figure 3. For ease of analysis let us assume that the reliability of each stage has the Bernoulli distribution with parameter Q and further that buffer i has capacity M_i. For a general N stage system with m buffers, there are $\prod_{i=1}^{m} (M_i + 1)$ possible buffer states; i.e., each buffer may take on $M_i + 1$ values and there are m such buffers. For example, if $M_i = 4$ for all i, and $m = 5$ there would be 3,125 possible buffer states ranging in value from (0,0,0,0,0) to (4,4,4,4,4). The question arises as to the viability of this form of analysis for systems with large buffer capacity (M_i), multiple buffers (m) or a combination of the two. Clearly, the transition matrix for a large system would be relatively sparse (i.e., many zero entries). For example, in a four stage (three buffer) system, where each buffer has a capacity of three units, there would be $4^3 = (3 + 1)^3$ or 64 possible transition states. For the starting state (1,1,1) there are 13 possible transitions (i.e., nonzero transition probabilities). The feasible transitions from the state (1,1,1) are:

(0,1,1), (0,1,2), (0,2,1), (1,0,1), (1,0,2), (1,1,0), (1,1,1),

(1,1,2), (1,2,0), (1,2,1), (2,0,1), (2,1,0), and (2,1,1).

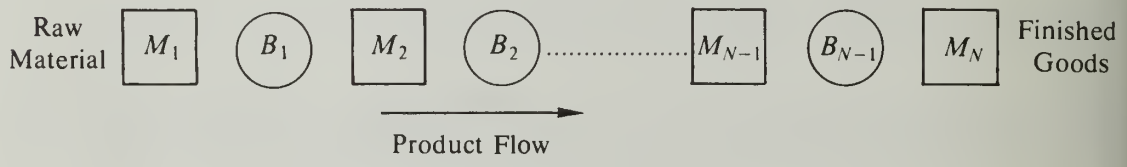


FIGURE 3. General multistage, multibuffer system

While it is obvious that the method of analysis employed to this point is feasible, that is, (1) definition of a one step transition matrix; (2) development of a reliability equation as a function of stage reliability and the steady state transition probabilities, and (3) solving a system of linear equations for the steady state transition probabilities; its application is, for the most part, not practical.

Let us present the transition matrices for two or three buffer systems with capacity one or two. For the system of two buffers of capacity two the transition matrix is given in Figure 4 and the steady state probabilities for various values of Q are given in Figure 5 where the reliability R is:

$$R = Q^3\pi_1 + Q\pi_2 + Q\pi_3 + Q^2\pi_4 + Q\pi_5 + Q\pi_6 + Q^2\pi_7 + Q\pi_8 + Q\pi_9.$$

Figure 5 was calculated by a small computer program. For various values of Q, we solved for the unique π_i and calculated R, which appears in Figure 5. For the system of four stages, and three buffers with capacity one, the transition matrix is given in Figure 6.

Again, using a small computer program we solved for π_i and calculated R. The steady state probabilities and the system reliability R is given in Figure 7 where

$$R = Q^4\pi_1 + Q\pi_2 + Q^3\pi_3 + Q\pi_4 + Q^3\pi_5 + Q\pi_6 + Q^2\pi_7 + Q\pi_8.$$

Transition Matrix

t \ t+1	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)
(0,0)	Q^3+P	Q^2P	0	QP	0	0	0	0	0
(0,1)	QP	Q^3+P^2	Q^2P	Q^2P	QP^2	0	0	0	0
(0,2)	0	QP	Q^3+P^2	0	Q^2P	QP	0	0	0
(1,0)	Q^2P	QP^2	0	Q^3+P^2	Q^2P	0	QP	0	0
(1,1)	0	Q^2P	QP^2	QP^2	Q^3+P^3	Q^2P	Q^2P	QP^2	0
(1,2)	0	0	Q^2P	0	QP^2	Q^3+P^2	0	Q^2P	QP
(2,0)	0	0	0	Q^2P	QP^2	0	Q^3+P	Q^2P	0
(2,1)	0	0	0	0	Q^2P	QP^2	QP	Q^3+P^2	Q^2P
(2,2)	0	0	0	0	0	Q^2P	0	QP	Q^3+P

FIGURE 4. Two buffers of maximum capacity two

Buffer Capacity Equals 2

	$Q = .9$	$Q = .7$	$Q = .3$	$Q = .1$
π_1	.108	.105	.097	.092
π_2	.092	.091	.090	.089
π_3	.060	.059	.060	.064
π_4	.126	.124	.118	.114
π_5	.106	.109	.121	.129
π_6	.092	.091	.090	.089
π_7	.183	.191	.209	.218
π_8	.126	.124	.118	.114
π_9	.108	.105	.097	.092
R	.855	.596	.205	.061

FIGURE 5. Exact solutions to three stage two buffer system

Maximum Buffer Capacity Equals One

t \ t+1	(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
(0,0,0)	Q^4+P	Q^3P	Q^2P	0	QP	0	0	0
(0,0,1)	QP	Q^4+P^2	Q^3P	Q^2P	Q^2P	QP^2	0	0
(0,1,0)	Q^2P	QP^2	Q^4+P^2	Q^3P	Q^3P	Q^2P^2	QP	0
(0,1,1)	0	Q^2P	QP^2	Q^4+P^2	0	Q^3P	Q^2P	QP
(1,0,0)	Q^3P	Q^2P^2	QP^2	0	Q^4+P	Q^3P	Q^2P	0
(1,0,1)	0	Q^3P	Q^2P^2	QP^2	QP	Q^4+P^2	Q^3P	Q^2P
(1,1,0)	0	0	Q^3P	Q^2P^2	Q^2P	QP^2	Q^4+P	Q^3P
(1,1,1)	0	0	0	Q^3P	0	Q^2P	QP	Q^4+P

FIGURE 6. Four stage, three buffer transition matrix

Buffer Capacity of One Unit

	$Q = .9$	$Q = .7$	$Q = .3$	$Q = .1$
π_1	.109	.100	.091	.080
π_2	.079	.075	.071	.073
π_3	.098	.098	.118	.138
π_4	.079	.080	.071	.021
π_5	.157	.156	.221	.270
π_6	.137	.147	.118	.193
π_7	.206	.213	.221	.187
π_8	.136	.131	.091	.038
R	.819	.533	.143	.036

FIGURE 7. Reliability of a four stage, three buffer system

The approach we present here can be summarized as follows; for a given configuration of a serial production line with multiple buffers and no restriction on their capacity, one can write the one step transition probability matrix and solve for its steady state probabilities which yields the reliability of the line. The method is efficient for a small number of buffers and small capacities. In general, the number of state variables and the number of linear equations are $\prod_{i=1}^{m} (M_i + 1)$ for m buffers with capacity M_i.

BIBLIOGRAPHY

[1] Buxey, G.M., N.D. Slack and R. Wild, "Production Flow Line Systems Design — A Review," AIIE Transactions (1973).

[2] Buxey, G.M. and D. Sadjadi, "Simulation Studies of Conveyor Paced Assembly Lines with Buffer Capacity," The International Journal of Production Research, *14* (1976).

[3] Buzacott, J.A., "Automatic Transfer Lines with Buffer Stocks," International Journal of Production Research, *5* (1967).

[4] Buzacott, J.A., "The Role of Inventory Banks in Flow-Line Production Systems," The International Journal of Production Research, *9* (1971).

[5] Buzacott, J.A., "Models of Automatic Transfer Lines with Inventory Banks, A Review and Comparison," AIIE Transactions, *10* (1978).

[6] Gershwin, S.B., "The Efficiency of Transfer Lines Consisting of Three Unreliable Machines and Finite Interstage Buffers," Presented at ORSA/TIMS Los Angeles Meeting (1978).

[7] Hatcher, J.M., "The Effect of Internal Storage on the Production Rate of a Series of Stages Having Exponential Service Times," AIIE Transactions, *2*, 150-156 (1969).

[8] Ignall, E. and A. Silver, "The Output of a Two-Stage System with Unreliable Machines and Limited Storage," AIIE Transactions, *9* (1977).

[9] Koenigsberg, E., "Production Lines and Internal Storage — A Review," Management Science, *5* (1959).

[10] Okamura, K. and H. Yamashina, "Analysis of the Effect of Buffer Storage Capacity in Transfer Line Systems," AIIE Transactions, *9* (1977).

[11] Rao, N.P., "Two Stage Production System with Intermediate Storage," AIIE Transactions, *7* (1975).

[12] Sheskin, T.J., "Allocation of Interstage Storage Along an Automatic Production Line," AIIE Transactions, *8* (1976).

[13] Soyster, A.L. and D.I. Toof, "Some Comparative and Design Aspects of Fixed Cycle Production Systems," Naval Research Logistics Quarterly, *23*, 437-454 (1976).

[14] Toof, D.I., "Output Maximization of a Series Assembly Facility Through the Optimal Placement of Buffer Capacity," Unpublished Ph.D. dissertation, Temple University (1978).

SCHEDULING COUPLED TASKS

Roy D. Shapiro

Harvard University
Graduate School of Business Administration
Cambridge, Massachusetts

ABSTRACT

Consider a set of task pairs coupled in time: a first (initial) and second (completion) tasks of known durations with a specified time between them. If the operator or machine performing these tasks is able to process only one at a time, scheduling is necessary to insure that no overlap occurs. This problem has a particular application to production scheduling, transportation, and radar operations (send-receive pulses are ideal examples of time-linked tasks requiring scheduling). This article discusses several candidate techniques for schedule determination, and these are evaluated in a specific radar scheduling application.

This article considers the problem of scheduling task pairs, i.e., tasks which consist of two coupled tasks, an initial task and a completion, separated by a known, fixed time interval. If the operator or machine performing these tasks is only able to process one at a time, scheduling is necessary to insure that a completion task of one pair does not arrive for processing while one part of another task is being processed.

Consider, for example, a radar tracking aircraft approaching a large airport [1]. In order to track adequately, it is necessary to transmit pulses and receive the reflection once every specified update period. The radar cannot transmit a pulse at the same time that a reflected pulse is arriving nor can two reflected pulses overlap. A possible strategy is to transmit to one tracked object and wait for that pulse to return before another pulse is transmitted as shown in Figure 1(a), but unless the number of objects being tracked is small, this may not allow all objects to be tracked in each update period. A more efficient strategy is some form of interlaced scheduling like that shown in Figure 1(b). Observe that the time between each pair of transmit and receive pulses is the same in Figure 1(b) as in Figure 1(a), yet the *total* transmission time is far less in 1(b).

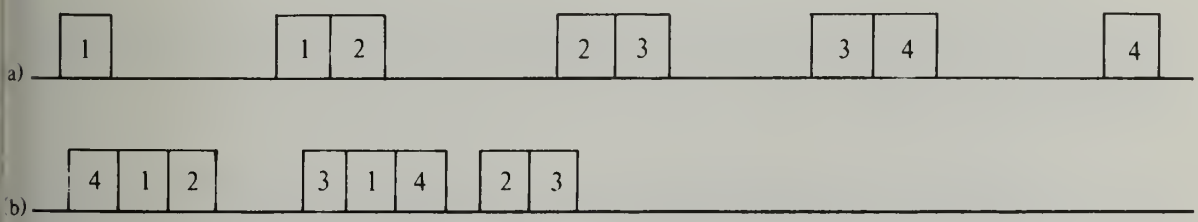


FIGURE 1. Sample 4-pair schedules

1. NOTATION, CLASSIFICATION, AND COMPLEXITY

Our object is to generate a schedule for a given set of task pairs which allows that set to be completed in the least possible time with no overlap between tasks (Figure 2). Formally, let

t_i = the time of initiation of the ith task pair;

S_i = the duration of the initial task of the ith pair, $i = 1, 2, \ldots, N$;

T_i = the duration of the completion task of the ith pair, $i = 1, 2, \ldots, N$;

d_i = the "inter-task" duration, i.e., the time between the *initiation* of the initial task of the ith pair and the *initiation* of that pair's completion.

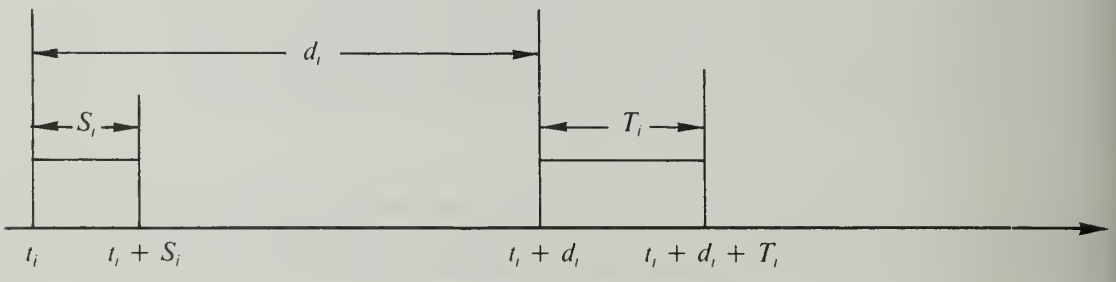


FIGURE 2. The ith task pair

The time between the initiation of the first task pair and the completion of the final pair we refer to as the *frame time* (or makespan, cf. [3,4] denoted z). For convenience, we will set the initiation time of the first pair to 0.

The scheduling problem may be stated as

$$\text{find } t_i \geqslant 0,\ i = 1, \ldots, N \text{ to minimize}$$

$$z = \max_i (t_i + d_i + T_i)$$

subject to the constraint that no member of the set of intervals

$$\{(t_i,\ t_i + S_i),\ (t_i + d_i,\ t_i + d_i + T_i)\}\quad i = 1, \ldots, N$$

overlap with any other member.

To put this problem into context with much of the recent literature classifying scheduling problems with regard to their computational complexity, we observe that the problem as stated is equivalent to a job shop problem where N jobs are to be scheduled on two machines with the following characteristics*:

1. Each job requires three operations: the first (of duration S_i) to be processed on Machine 1; the second (of duration $d_i - S_i$) on Machine 2; the third (of duration T_i) again on Machine 1.

2. Machine 1 may only process one operation at a time; Machine 2, however, has infinite processing capacity.

*Under the classification scheme of Rinooy Kan [9], this problem is $V|2|$ *G, no wait, M_2 non-bott* $|C_{\max}$. See also [8].

3. No waiting between operations is permitted. That is, once a job is begun, it must proceed from Machine 1 to Machine 2 and back again to Machine 1 with no delay.

The problem can then be shown to be NP-complete by Theorem 5.7, pg. 93 in [9] or by a reduction from KNAPSACK in [6]. NP-complete problems form an equivalence class of combinatorial problems for which no nonenumerative algorithms are known. If an "efficient" algorithm were constructed which could solve any problem in this class, any other would also be solvable in polynomial time (cf. [2,4,6,7]). Members of this class include the chromatic number problem, the knapsack problem, and the traveling salesman problem.

The fact that a polynomial-bounded algorithm is not likely to exist motivates the construction of several polynomial-bounded algorithms which are presented and evaluated in Sections 2 and 3. An integer programming formulation leads to a straightforward branch and bound procedure which makes use of the problem's special structure. (See [11].) In view of the fact that this optimal procedure is likely to be tractable only for very small problems, and not even then for radar-like applications requiring real time solution, we proceed directly to consideration of three suboptimal algorithms.

2. SUB-OPTIMAL ALGORITHMS

This section considers scheduling procedures which can be shown to be polynomially bounded: Sequencing, Nesting and Fitting. After some discussion of their characteristics, they will be evaluated on realistic examples in Section 3.

Sequencing

An ordered set of p task pairs are said to be sequenced when the completion tasks arrive for processing in the same order as the initial tasks were scheduled. p pairs can be sequenced whenever

$$d_1 \geqslant \sum_{i=1}^{p} S_i \quad \text{and} \tag{1}$$

$$d_i \geqslant d_{i-1} + T_{i-1} - S_{i-1}, \quad i = 2, 3, \ldots, p. \tag{2}$$

If, as is the case for many applications, $S_i = T_i$ for each task pair, (2) becomes simply

$$d_i \geqslant d_{i-1}, \tag{3}$$

and implementation of this procedure becomes quite easy.

We may think of this procedure as "jamming" initial tasks together until they run into the completion task corresponding to the first initial task. The completion tasks are guaranteed not to overlap since each succeeding d_i is at least as large as the one before. Also, since this is a "single-pass" procedure (cf. [3]), computation time is linear in N.*

In any sequenced p-set, dead time can occur in two ways, as is shown in Figure 3. It occurs between the last initial task and the first completion, and it occurs between successive completions. The former can be written as

$$d_1 - \sum_{i=1}^{p} S_i$$

*Actually, computation time is $O(N \log N)$ since the d_i have to be ordered.

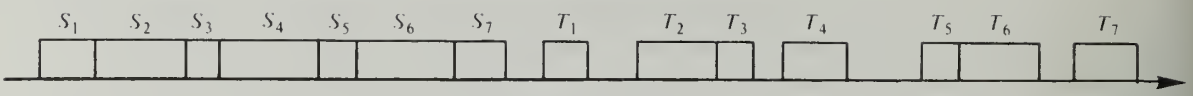


FIGURE 3. Sequencing

and the latter as

$$\sum_{i=1}^{p} d_{i+1} - d_i = d_p - d_1.$$

Hence,

$$z_{SEQ} = \sum_{i=1}^{p} (S_i + T_i) + \left[d_1 - \sum_{i=1}^{p} S_i \right] + d_p - d_1$$

$$= \sum_{i=1}^{p} T_i + d_p.$$

Hence, if N task pairs are sequenced in P p-sets, the kth set having p_k pairs, $k = 1, 2, \ldots, P$, the total frame time may be represented as

$$z_{SEQ} = \sum_{k=1}^{P} \left[\sum_{i=1}^{p_k} T_i + d_{p_k} \right] = \sum_{i=1}^{N} T_i + \sum_{k=1}^{P} d_{p_k}.$$

As an example, consider the following 7 task pairs with common durations for initial and completion tasks, ordered by increasing d_i.

$i = 1$: $S_1 = T_1 = 2$, $d_1 = 9$

$i = 2$: $S_2 = T_2 = 1$, $d_2 = 13$

$i = 3$: $S_3 = T_3 = 2$, $d_3 = 15$

$i = 4$: $S_4 = T_4 = 3$, $d_4 = 15$

$i = 5$: $S_5 = T_5 = 2$, $d_5 = 19$

$i = 6$: $S_6 = T_6 = 4$, $d_6 = 24$

$i = 7$: $S_7 = T_7 = 3$, $d_7 = 25$.

Figure 4(a) shows their sequenced schedule.

For comparison, Figure 4(b) shows the optimal schedule for this set of task pairs as generated by the branching algorithm alluded to above. At the other extreme, if these pairs were scheduled by waiting until each pair was completely processed before initiating the next, the frame time would be 138.

$$\left[z = \sum_{i=1}^{7} (d_i + T_i) = 138. \right]$$

Nesting

An ordered set of p task pairs are said to be nested whenever the completion tasks arrive for processing in the reverse of the order in which the initial tasks were scheduled. p pairs may be nested if

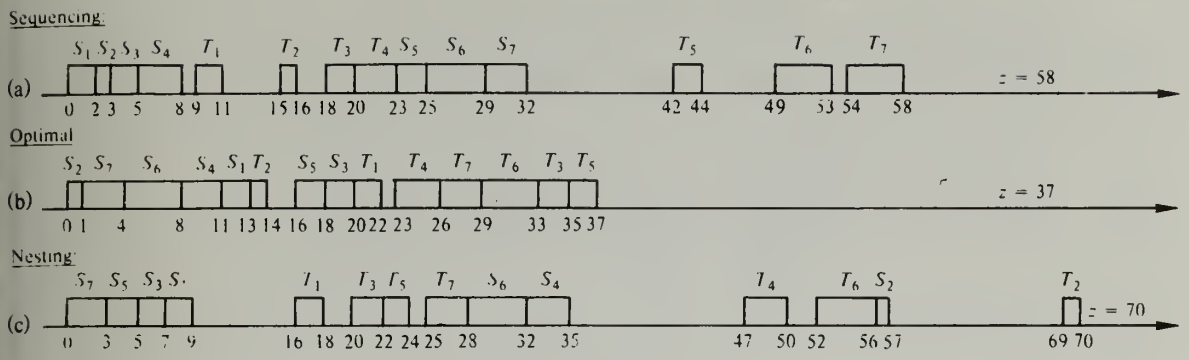


FIGURE 4. Sequencing and nesting

(4) $$d_i \geqslant d_{i+1} + T_{i+1} + S_i, \quad i=1, \ldots, p-1.$$

Applying this procedure to the 7-pair example discussed above gives the schedule shown in Figure 4(c) with $z = 70$.

Fitting

This procedure, unlike the two discussed above, allows the user to specify a priority ordering, and corresponds intuitively to the simple process which one might use when scheduling task pairs by hand. After setting the desired order and scheduling the first task pair at time 0, each successive pair is scheduled at the earliest possible time not involving any overlap with pairs already scheduled.

Let us consider this procedure for the above example, taking an arbitrary ordering: 2,6,7,4,3,1,5. As shown in Figure 5(a), the task pair is scheduled at time 0, and pairs 6 and 7 can successively be scheduled with no overlap. If we, however, try to schedule pair 4 at the first available time, its completion would overlap with pair 6's completion (Figure 5(b)), so this is not possible. The first available time for scheduling task pair 4 without overlap is time 18 (Figure 5 (c)). Pair 3, however, having task duration only 2, *can* be scheduled at time 8 (Figure 5(d)). Observe now that pair 1 can be scheduled nowhere in the existing schedule without overlap, so it must be "tacked" onto the end, at time 36 (Figure 5(e)). Pair 5 is scheduled at time 21, completing the schedule with $z = 47$ (Figure 5(f)).

3. TASK PAIR SIMULATION AND NUMERICAL RESULTS

In keeping with the radar application mentioned above, a simulation has been developed to generate aircraft configurations suitable to radar operation. For each object, range, cross-section, and velocity can be used to determine the necessary length of transmit and receive pulses (of the order of 10-100 μsecs.) as well as the inter-pulse distance (of the order of 300-1300 μsecs.). Thus, a list of task pairs can be generated for evaluation of the procedures outlined in the previous section. As an example, such a list is given in Table I for $N = 20$.

For values of N shown in Table II, the simulation generated 50 such task pair lists, and the average frame time and computation time were computed. Figure 6 presents this data graphically. Note that, as one would expect, frame time is linear in N. This is not surprising since in the best conceivable situation, that of no idle time between subtasks,

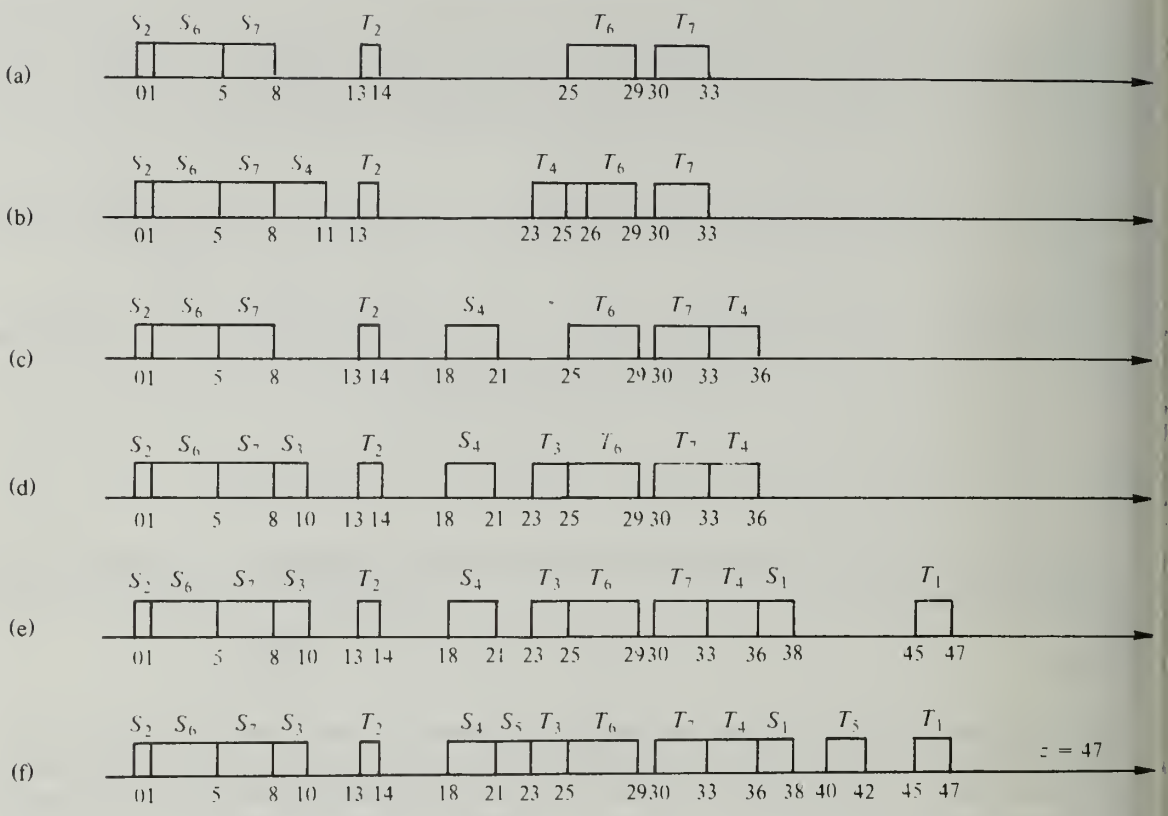


FIGURE 5. Fitting

TABLE I — *Sample Task Pair List (N = 20)*

i =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$S_i = T_i$ (μsec)	70	70	80	50	70	80	70	70	70	40	50	70	60	60	60	70	50	40	50	40
d_i (μsec)	1334	1258	1254	1159	1107	1022	954	884	791	750	709	674	631	623	621	555	513	498	465	387

TABLE II (a) — *Average Simulated Frame Times*

N	SEQUENCE	NEST	FIT
20	4.4 (.414)	7.3 (1.387)	4.0 (.381)
50	8.6 (.559)	15.0 (1.830)	7.6 (.452)
100	15.5 (.759)	27.2 (2.747)	13.8 (.679)
200	29.2 (1.197)	52.2 (4.683)	27.4 (.990)
500	70.8 (2.188)	119.2 (8.391)	66.1 (1.590)

Frame times in msec.
Quantities in parentheses are standard deviations

TABLE II (b) — *Average Computation Times (msec)*

N	SEQUENCE	NEST	FIT
20	1.9	4.0	67.2
50	4.4	16.3	440.1
100	8.6	51.9	1742
200	16.9	195.3	7091
500	42.0	1064	44160

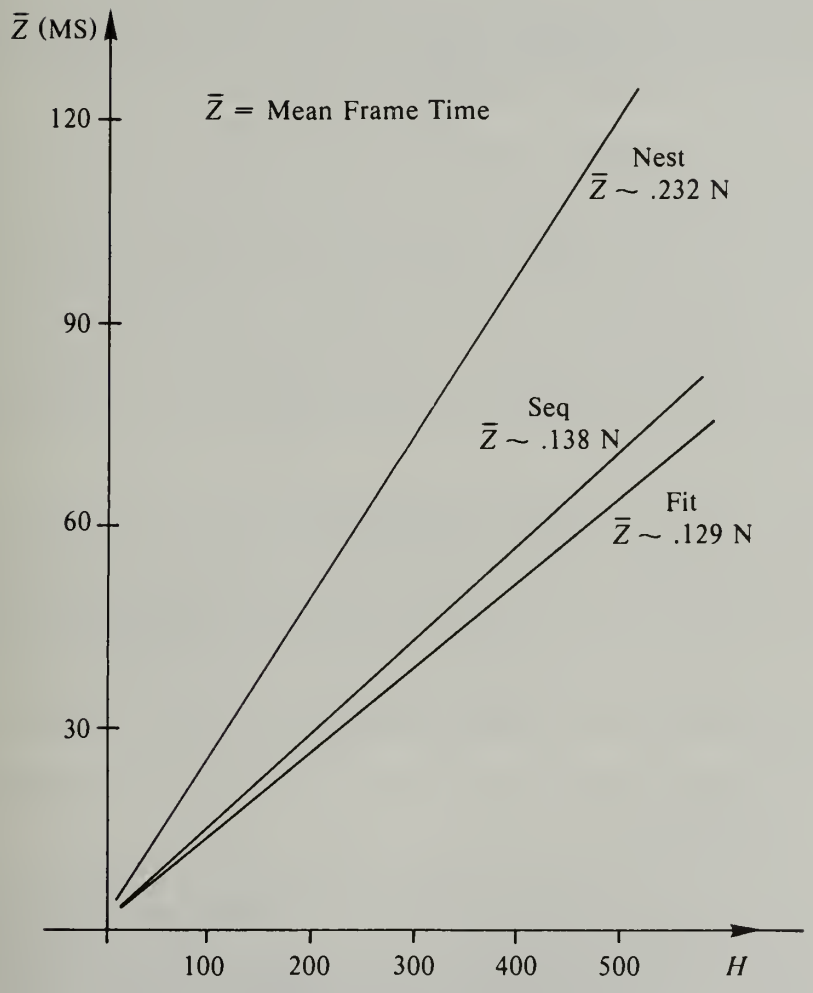


FIGURE 6. Comparative frame times

$$z = \sum_{i=1}^{N} (S_i + T_i) \sim k_1 N$$

and in the worst situation, that in which no task is performed until the previous task has been completed,

$$z = \sum_{i=1}^{N} d_i + T_i \sim k_2 N.$$

An assumption made in the treatment of this example is that the radar operator knows the values of S_i, T_i and d_i precisely. If there is any uncertainty, signals can overlap. A straightforward way of avoiding this problem in a real situation where uncertainty would obviously be present would be to "open a window" around the pulse. That is, if the object is such that transmit and receive pulses are estimated to be of 60 μsec. duration, an interval longer than 60 μsec. can be allotted to these pulses to accommodate (1) the possibility that a pulse length longer than 60 μsec. might be necessary or, more important, (2) the possibility that the receive signal might arrive sooner or later than expected. This procedure offers no conceptual difficulty since the window around the pulses may be made large enough to guarantee that the probability of overlap is as small as required. In order to retain frame times small enough to allow updating every, say, 200 milliseconds, we must limit the size of the window somewhat. This does not seem to be a severe restriction, however. For example, since frame time is linear in ΣT_i, opening a window around each pulse of twice that pulse's estimated duration would cause the frame time to be no more than doubled. The frame times of sequenced pulses in Table II(a) indicate that even for large N, this is no problem.

A second possibly problematic characteristic of the example is that it is static, i.e., no explicit consideration is given to new "jobs" added to the system during the scheduling process. In job shop scheduling, this may present no problem if jobs are released to the shop at predetermined times. In radar tracking, however, one cannot hold enemy missiles, and the scheduler must be dynamic. This can be accomplished; the new targets may be inserted into the queue of jobs to be processed, or, since this is likely to be time-consuming when jobs are ordered (as in Sequencing and Nesting), all current jobs can be processed, followed by the newly-arrived entries. This procedure will be especially efficient for sequencing since the d_i's are proportional to the distance between radar and target, and new targets will tend to appear at approximately the same range.

The necessity to allow for search and discrimination as well as the tracking activity and real-time schedule determination within a 200 milli-second period makes sequencing the only viable alternative. Even when real-time processing is not required, one wonders whether the slight improvement in frame time allowed by fitting warrants the extra computational burden.

A *caveat* is in order here: these results are somewhat application-dependent. It is quite possible that other applications which produce task pairs with different structures will lead to different conclusions.

CONCLUDING REMARKS

In the above discussion it has been assumed that the operator or machine can process only one task segment at a time. This is appropriate for the application being considered, but one might easily imagine instances in which there is some nonunit capacity constraint on the operator. For example, if trucks are being loaded and unloaded at some central depot, labor or space restrictions might limit the number of trucks being simultaneously processed.

Fortunately, the suboptimal procedures described above may be extended without any problem.* Figure 7 shows how the example given in Section 2 may be sequenced if the operator is limited to two tasks at a time. Note that due to the ordering of the inter-task durations, sequencing guaranteed that since no more than two initial tasks can overlap, no more than two final tasks will overlap.

*The optimal enumerative procedure described in [11] is also easily extended.

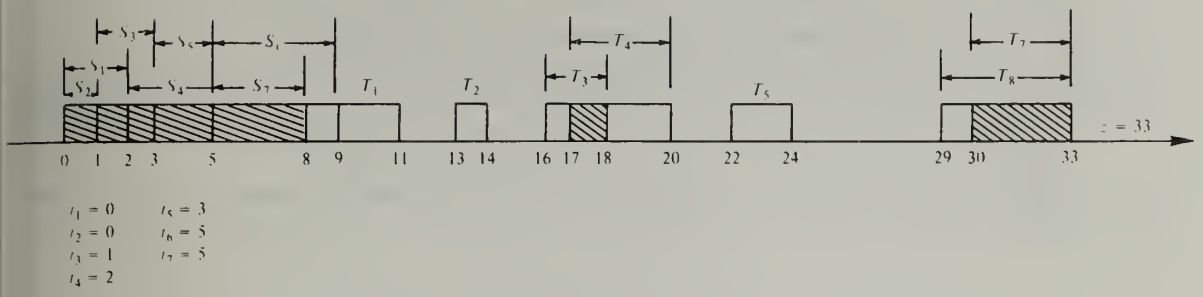


FIGURE 7. Sequencing with operator capacity = 2

Another extension is to consider tasks which consist of more than two coupled segments. The notation changes slightly: the ith task pair becomes a task set, the initial task of duration S_i followed by n_i subtasks; the jth subtask is of duration T_{ij} and the time at which it is initiated is d_{ij} after the initiation of the initial task (Figure 8).

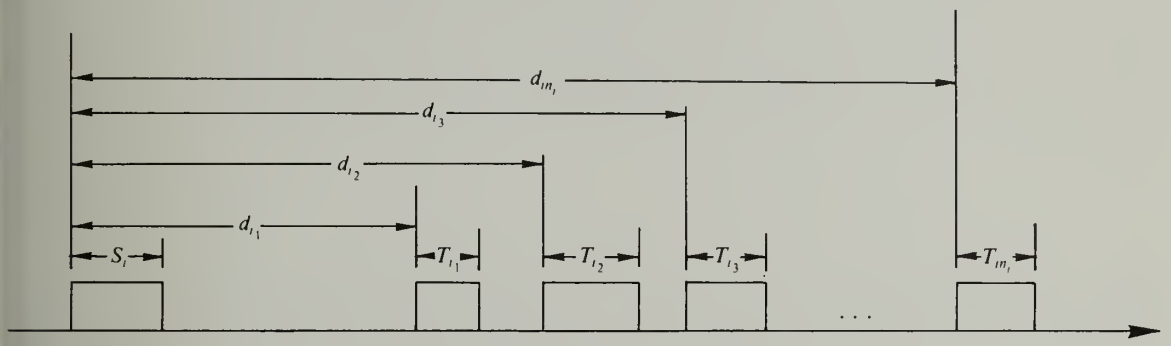


FIGURE 8. Multiple coupled subtasks

Fitting, as proposed above, works well in this case, but sequencing and nesting are wasteful since they treat the subtasks as one long task of duration $d_{in_i} + T_{in_i} - d_{i1}$.

BIBLIOGRAPHY

[1] Air Traffic Control Advisory Committee, Report of the Department of Transportation, *1* (1969).

[2] Coffman, E.G., Jr. (editor), *Computer and Jobshop Scheduling Theory*, John Wiley, New York, N.Y. (1976).

[3] Conway, R.W., W.L. Maxwell and L.W. Miller, *Theory of Scheduling*, Addison-Wesley, Reading, Mass. (1967).

[4] Garey, M.R., D.S. Johnson and R. Sethi, "The Complexity of Flowshop and Jobshop Scheduling," *Mathematics of Operations Research, 1*, 117-129 (1976).

[5] Heffes, J. and S. Horing, "Optimal Allocation of Tracking Pulses for an Array Radar," *IEEE Transactions on Automatic Control, 15*, 81-87 (1970).

[6] Karp, R.M., "Reducibility among Combinatorial Problems," R.E. Miller and J.W. Thatcher, (editors), *Complexity of Computer Computations*, Plenum Press, New York, N.Y., 85-104 (1972).

[7] Karp, R.M., "On the Computational Complexity of Combinatorial Problems," *Networks, 5*, 45-68 (1975).

[8] Reddi, S.S. and C.V. Ramamoorthy, "On the Flowshop Sequencing Problem with No Wait in Process," *Operational Research Quarterly, 23*, 323-331 (1972).

[9] Rinooy Kan, A.H.G., *Machine Scheduling Problems*, Martinus Nijhoff, The Hague (1976).

[10] Schweppe, F.C. and D.L. Gray, "Radar Signal Design Subject to Simultaneous Peak and Average Power Constraints," *IEEE Transactions on Information Theory, 12*, 13-26 (1966).

[11] Shapiro, R.D., "Scheduling Coupled Tasks," Harvard Business School, Working Paper, HBS 76-10.

SEQUENCING INDEPENDENT JOBS WITH A SINGLE RESOURCE

Kenneth R. Baker

Dartmouth College
Hanover, New Hampshire

Henry L. W. Nuttle

North Carolina State University
Raleigh, North Carolina

ABSTRACT

This paper examines problems of sequencing n jobs for processing by a single resource to minimize a function of job completion times, when the availability of the resource varies over time. A number of well-known results for single-machine problems which can be applied with little or no modification to the corresponding variable-resource problems are given. However, it is shown that the problem of minimizing the weighted sum of completion times provides an exception.

1. INTRODUCTION

We consider the problem of sequencing a set $N = \{1, 2, \ldots, n\}$ of jobs to be processed using a single homogeneous resource, where the availability of the resource varies over time. If t represents time (measured from some origin $t = 0$) then we denote by $r(t)$ the resource available at time t and by $R(t)$,

$$R(t) = \int_0^t r(u)\,du$$

the cumulative availability as of time t, i.e., the area under the curve $r(u)$ over the interval $[0,t]$. See Figure 1.

Let p_j, $j = 1, \ldots, n$, denote the resource requirement of job j. Once p_j units of resource have been applied to job j, the job is considered complete. We denote the completion time of job j by C_j. In all problems treated the objective is is to minimize G, a function of the completion times of the jobs, where G is assumed to be a regular measure (see [1], Chapter 2).

This model is a generalization of the single-machine sequencing model. The generalization to a resource capacity that varies over time allows for situations in which machine availability is interrupted for scheduled maintenance or temporarily reduced to conserve energy. It also allows for a situation in which processing requirements are stated in terms of man-hours and labor availability varies over time.

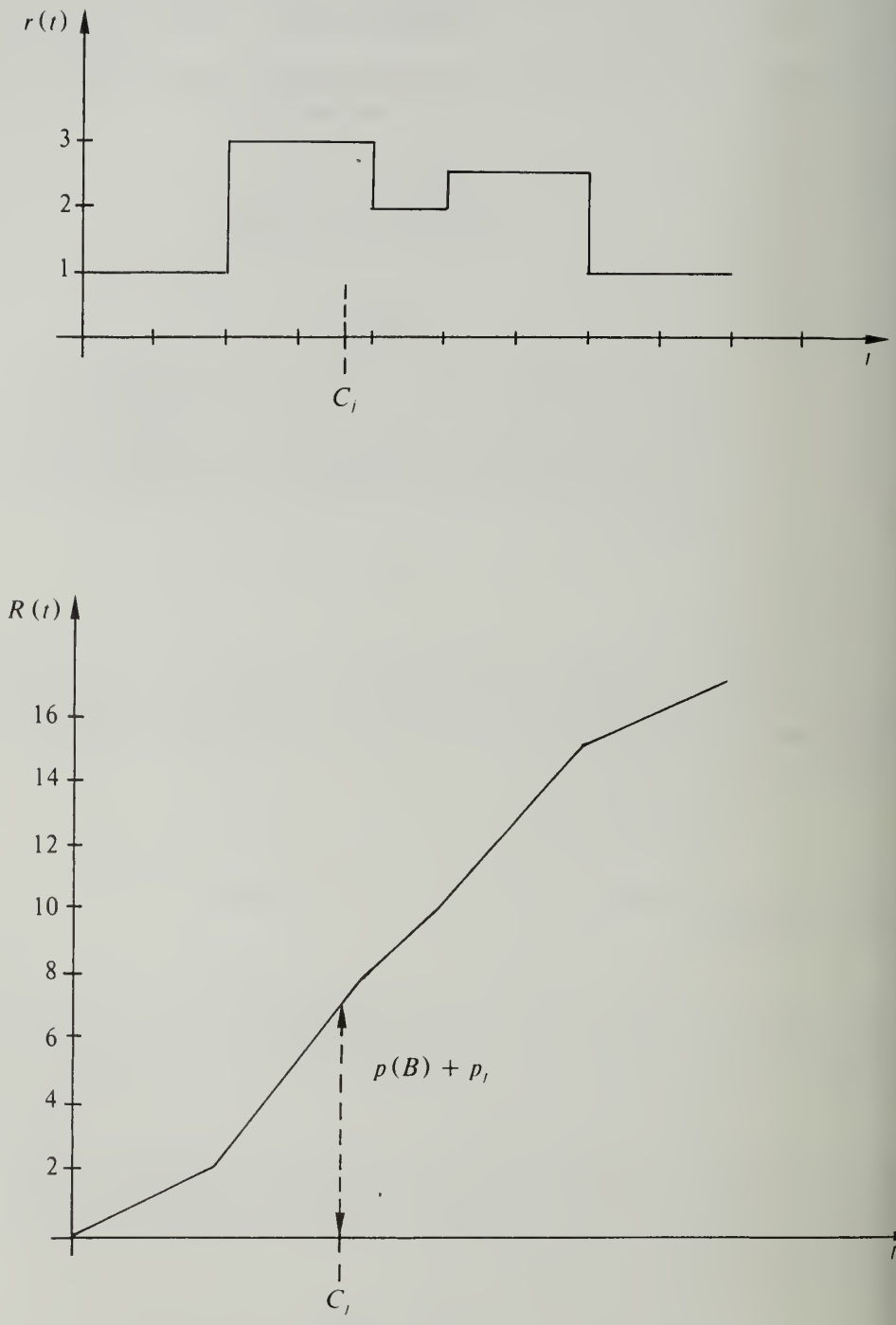


FIGURE 1.

In the single-machine case the resource profile $r(t)$ is constant (typically $r(t) = 1$), and the cumulative profile $R(t)$ is a straight line with slope $r(t)$. Time is measured in some basic unit such as hours; and completion times, ready times, due dates and tardiness are expressed in the same units. Resource requirements (processing times) are simply requirements for intervals on the time-axis.

In the variable-resource problem, the exact correspondence between the requirement for a unit of resource and the requirement for a unit interval on the time axis is lost. This lack of correspondence arises from the fact that there may be a number of units of resource available during a particular unit of time and a different number during the next. In the single-machine problem if a job j is sequenced to follow jobs in B (where B is any subset of N) then job j will be complete at time C_j,

$$C_j = p(B) + p_j,$$

where $p(B) = \sum_{i \in B} p_i$, and p_i denotes the processing time for job i. In the variable-resource problem it is appealing to analogously specify the completion time of job j by C_j,

$$C_j = t(p(B) + p_j) \tag{1}$$

where p_i is the resource requirement of job i and $t(Q)$ is the (smallest) point on the time axis corresponding to $R(t) = Q$. See Figure 1. In effect, jobs are sequenced on the resource axis, while their completion times are measured on the time axis. For the single-machine problem the completion point of job j is the same on both axes, but such is not the case for the variable-resource problem.

Notice that this specification implicity assumes that the resource available at any point in time is devoted entirely to the processing of a single job. Thus, for example, if ten men were available in a particular hour, all ten would be assigned to work simultaneously on the same job. Also, if the available resource represents several machines, then this formulation permits each job to be processed simultaneously on more than one machine. Equivalently, this means that jobs must be divisible into portions that can be allocated equally to the number of machines available. Such a formulation will be called a continuous-time model.

In order to allow for a wider range of applicability, we can re-formulate the model in discrete time as follows.

(a) Unit intervals on the time axis (of Figure 1) are called periods, and job completion times are measured in periods.

(b) In a given period the resource availability is an integer number of units.

(c) Each job requires an integer number of resource-periods.

(d) Processing work is divisible only to the level of one resource-unit for one period.

Under this formulation, for example, the time unit might be days, the resource availability might be crew size, and the processing requirement might be man-days. Property (d) then restricts the refinement of a schedule to the assignment of each crew member's task on a day-by-day basis. Furthermore, a task requiring two man-days could be accomplished either by one crew member working two days or by two members working one day each.

In the discrete-time context, we may regard sequencing as ordering jobs on the resource scale in Figure 1, but taking the completion time of job j to be $[C_j]$ vs $[C_j]$, the smallest

integer greater than or equal to C_j, where C_j itself is given by (1). In other words, we obtain a sequence using the continuous-time framework, which assumes arbitrarily divisible jobs, but we round up the resulting completion times when they are noninteger. Under this interpretation of the model, due-dates are specific days and a job is "on time" as long as it is completed on or before the specific day. Clearly, in the discrete time model several jobs can have the same completion time.

To verify that a job sequence can be interpreted consistently with requirement (d), note that the cumulative resource requirement and the cumulative resource availability by the end of any period are both integers. It follows that the workload implied by the continuous-time solution can be shifted to meet the integer restrictions of the discrete-time model since the resource availability in any period can be treated as a set of unit-resource availabilities. Then any fraction of a day's work in the original solution can be rescheduled as a day's work for the same proportion of the total resource units available. This rescheduling will consume an integer number of resource-periods for each job.

As an example, consider the three-job problem shown below.

j	1	2	3
p_j	7	3	6

$$r(t) = 1 \quad 0 \leqslant t \leqslant 4$$
$$r(t) = 4 \quad 4 \leqslant t \leqslant 7$$

In Figure 2 we represent the sequence 1-2-3 assuming infinite divisibility. In Figure 3 we show how the work is rescheduled to meet the integrality requirement of the discrete-time model. As Figures 2 and 3 indicate, the discrete-time conditions can be incorporated by a minor adjustment of continuous-time job assignments that essentially involves replacing vertical portions of the schedule chart with horizontal portions whenever the available resource capacity is split among two or more jobs within a period.

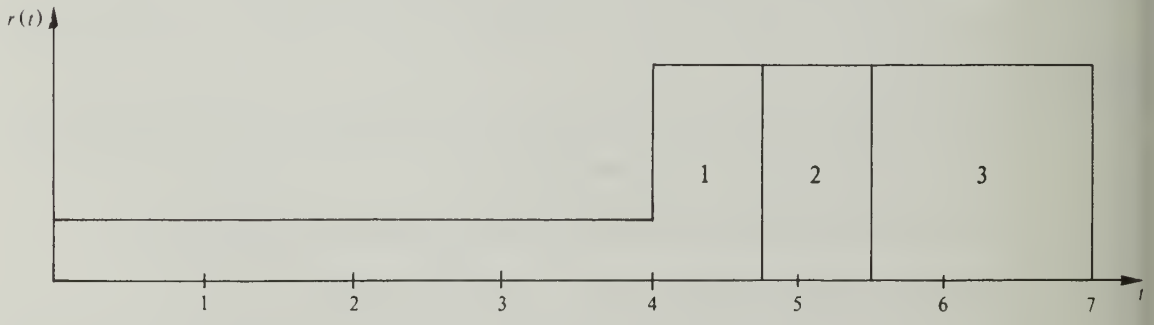


FIGURE 2.

Our purpose in this paper is to note that certain well-known results for the single-machine model carry over with little or no modification to the variable-resource model. In fact, we found only one exception. (See Section 3.)

A variable resource problem has also been examined by Gelders and Kleindorfer [6,7] in the context of coordinating aggregate and detailed scheduling decisions. In their model the variation in resource availability results from the explicit decision to schedule overtime. This

decision leads to a cumulative resource availability function consisting of segments with identical positive slope (corresponding to capacity available) separated by horizontal segments (corresponding to unused overtime.) Their objective is to determine when and how much overtime should be scheduled, and to determine the associated job sequence, so as to minimize the sum of overtime, tardiness and flow-time costs. They also note that for a given overtime schedule, shortest-first sequencing minimizes mean job completion time while nondecreasing processing time-to-weight ratio sequencing may not minimize mean weighted job completion time. These two results are encompassed in our general treatment of the variable-resource model in Sections 2 and 3.

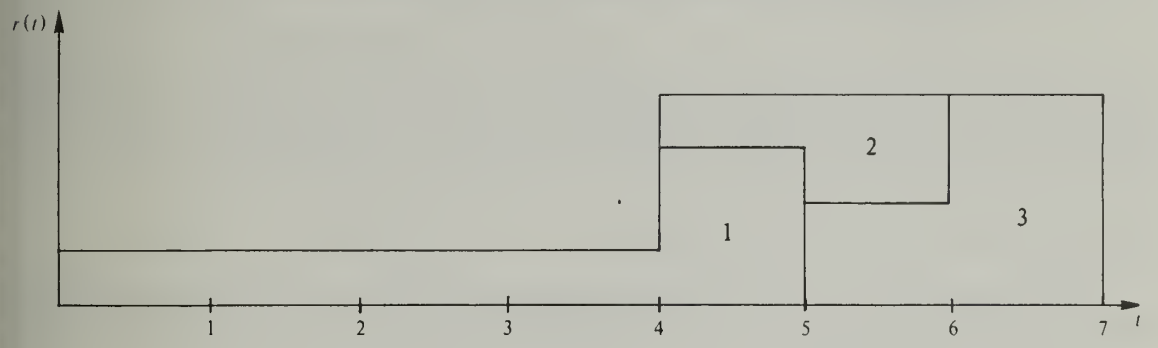


FIGURE 3.

2. RESULTS THAT GENERALIZE TO VARIABLE RESOURCES

The following is a set of sequencing results for the variable-resource model that are identical to or slight modifications of their single-machine counterparts. It is not difficult to establish that the results we give are valid for both the continuous-time and discrete-time models. However, proofs are omitted, since they are typically direct extensions of the original arguments in the single-machine case.

The results involve sequences of jobs, or at least partial sequences. We reiterate that these sequences can be viewed as applying to the resource axis in Figure 1 but can be converted to completion time schedules in either the continuous-time or discrete-time case by means of the appropriate transformation. We use C_j to denote the completion time of job j and $t(p(B))$ to denote the makespan for the jobs in B, recognizing that in the discrete-time case these quantities must be interpreted in the appropriate way.

Minimizing the Maximum Cost

One of the few efficient algorithms for a broad class of sequencing criteria is Lawler's procedure [9] for minimizing the maximum cost in the sequence. Formally, the criterion is to minimize

$$G = \max \{g_j(C_j)\}$$

where $g_j(C_j)$ is the cost incurred by job j when it completes at C_j and where $g_j(t)$ is nondecreasing in t. The solution procedure works by constructing a sequence from the back of the schedule and the procedure is easily adapted to the variable-resource model, as shown below.

1. Initially let $A = \phi$. (A denotes the set of jobs at the end of the schedule and $A' = N - A$ denotes its complement.)

2. Find $M = t(p(A'))$. (M is the makespan for unscheduled jobs.)

3. Identify job k satisfying $g_k(M) = \min_{j \in A'} \{g_j(M)\}$. (Considering only the unscheduled jobs, job k is the one that achieves the minimum cost when scheduled last.)

4. Schedule job k last among the jobs in A'. Then add job k to A and return to Step 2 until $A = N$.

A noteworthy special case is the criterion of maximum tardiness. The procedure sequences jobs in nondecreasing order of due-dates in this case. Thus, as in the single machine problem, earliest due-date (EDD) scheduling will minimize the maximum tardiness. It will also find a schedule in which all jobs complete on time, if such a schedule exists.

Minimizing the Sum of Tardiness Penalties

Many problems of considerable interest for the single machine model may be regarded as special cases of the problem of minimizing total tardiness penalty,

$$G = \sum_{j \in N} w_j T_j$$

where $T_j = \max(C_j - d_j; 0)$ and $w_j > 0$.

Several dominance properties, in the spirit of Emmons [3], can be shown to hold for the variable resource problem. These in turn imply similar dominance properties for the various special cases and, in some instances, optimizing (ranking) procedures. Let:

J = a set of jobs
J' = the complement of J
A_i = the set of jobs known to follow job i, by virtue of precedence conditions.
B_i = the set of jobs known to precede job i, by virtue of precedence conditions.
C_J = the time required to process the jobs in set J, defined by $R(C_J) = \sum_{j \in J} p_j$
B_{ij}^{*} = $B_i \cup \{j\}$ = the set containing job j and all jobs known to precede job i by virtue of the precedence conditions.
A_{ij}^{*} = $A_i' - \{j\}$ = the set containing the complement of A_i, but excluding job j.

THEOREM 1: If $w_k \leqslant w_j$ and $d_k \geqslant \max(d_j, C_{A_{jk}^{*}})$ then j precedes k in an optimal sequence.

THEOREM 2: If $d_k \geqslant C_{A_j'}$ then j precedes k in an optimal sequence.

THEOREM 3: If $p_j \leqslant p_k$, $w_j \geqslant w_k$ and $d_j \leqslant \max(d_k, C_{B_{kk}^{*}})$, then j precedes k in an optimal sequence.

COROLLARY (Theorem 3): If $w_j \geqslant w_k$, $p_j \leqslant p_k$ and $d_j \leqslant d_k$, then j precedes k in an optimal sequence. The corollary immediately yields an optimal ranking procedure for problems derived by making constant any two of the three parameters. For example, when $G = \sum_{j \in N} T_j$ with $w_j = w$ and $d_j = d$, an optimal sequence is determined by ordering the jobs by processing

requirement, smallest first ($p_1 \leqslant p_2 \ldots \leqslant p_n$). When $d = 0$ we have $T_j = C_j$, i.e., the mean flowtime problem, for which this sequence is called shortest processing time (SPT).

The problem of minimizing the total tardiness penalty when $p_j = p$ is also not difficult to solve. Constant resource requirements imply a fixed sequence of completion times under any sequence. In particular the first job completes at $t(p)$, the second job at $t(2p)$, etc.; and an optimal schedule may be found by assigning jobs to positions, as in Lawler [10]:

x_{ij} = 1 if job i appears in sequence position j
= 0 otherwise
c_{ij} = the penalty for job i when it appears in sequence position j, i.e. max $\{0, t(jp) - d_i\}$.

The problem is to minimize $\sum_i \sum_j c_{ij}x_{ij}$

Subject to

$$\sum_i x_{ij} = 1$$

$$\sum_i x_{ij} = 1.$$

An assignment algorithm can produce the optimal solution.

The most general version of the single-machine problem, with unequal due-dates, processing times, and weights is binary NP-complete. The computational complexity of the cases in which $w_i = w$ or $d_i = d > 0$ is an open question. However, pseudo-polynomial algorithms have been developed by Lawler [11] and Lawler and Moore [12]. The algorithms which have demonstrated the most effective computational power for the problems are those found in [14]. These and other enumerative algorithms can be modified in a straightforward manner to accommodate the variable resource problem.

Minimizing The Weighted Number of Tardy Jobs

In this case we are interested in whether a job is tardy rather than the the length of time by which it is tardy. Let $\delta(T_j) = 1$ indicate that job j is tardy and $\delta(T_j) = 0$ indicate that it is completed on time. If each job has its own penalty for being tardy, i.e.,

$$G = \sum_{j \in N} w_j \delta(T_j),$$

then the single-machine problem is binary NP-complete, although it can be solved by a pseudo-polynomial dynamic programming algorithm due to Lawler and Moore [12]. The algorithm can easily be adapted to the variable-resource problem with no impact on computational efficiency.

By restricting the data we obtain special cases that are solvable by ranking algorithms, just as in the single-machine case:

THEOREM 4: When $d_j = d$ for all jobs, if the processing times and weights are agreeable ($p_j \leqslant p_i$ whenever $w_j \geqslant w_i$) then an optimal sequence is obtained by scheduling the jobs in order of processing requirement, shortest first (in order of weight, largest first).

COROLLARY (Theorem 4): When $d_j = d$ and $p_j = p$, an optimal sequence is obtained by scheduling jobs in order of weight, largest first.

When $w_j = w$ for all jobs a sequence that minimizes the number of tardy jobs i.e., $G = \sum_{j \in N} \delta(T_j)$, can be determined by generalizing an efficient algorithm due to Moore [13]. Since maximum tardiness is minimized by sequencing the jobs in EDD order, it follows that if sequence S yields minimum G, then so will sequence S', in which the on-time jobs in S are scheduled in EDD order followed by all the tardy jobs in S. Letting S_n represent the largest possible set of on-time jobs (so that $G = n - |S_n|$ is the minimum number of tardy jobs) S_n can be determined as follows:

1. Order and index the jobs in N such that $d_1 \leqslant d_2 \leqslant \ldots \leqslant d_n$ (where ties are broken arbitrarily). Set $S_0 = \emptyset$ and $k = 1$.

2. If $k = n + 1$ stop. S_n is an optimal set.

3. If $t\left(\sum_{i \in S_{k-1}} p_i + p_k\right) \leqslant d_k$ set $S_k = S_{k-1} \cup \{k\}$, otherwise let $p_r = \max \{p_j | j \in S_{k-1} \cup \{k\}\}$ and set $S_k = S_{k-1} \cup \{k\} - \{r\}$.

4. Set $k = k + 1$ and return to step 2.

Constrained (Secondary Criterion) Problems

Several authors have addressed the problem of sequencing n jobs on one machine so as to optimize one criterion while restricting the set of sequences so that all or some jobs also satisfy another. We include four such problems here. In particular,

(a) Minimize total (mean) flow time given that a subset E of the jobs are to be on time (Burns and Noble [2] and Emmons [4], i.e.,

$$\min G = \sum_{i \in N} C_i$$

s.t. $C_i \leqslant d_i$, $i \in E$

(b) Minimize maximum tardiness given that a subset E of the jobs are to be on time (Burns and Noble [2]), i.e.,

$$\min G = \max_{i \in N} T_i$$

s.t. $C_i \leqslant d_i$, $i \in E$

(c) Minimize mean flow time over all sequences which yield minimum maximum cost (Emmons [5] and Heck and Roberts [8]), i.e.,

$$\min G = \sum_{i \in N} C_i$$

s.t. $g_i(C_i) \leqslant G_m$, $i \in N$

where $g_i(C)$ is a non-decreasing function of C and $G_m = \min \{\max g_i(C_i)\}$

(d) Minimize the number of tardy jobs given that a subset E of the jobs is to be on time (Sidney [15]), i.e.,

$$\min G = \sum_{i \in N} \delta(T_i)$$

subject to $C_i \leqslant d_i$ $i \in E$.

In all cases the algorithms originally developed for the single-machine problem can easily be adapted to the variable-resource problem.

The first three problems can be solved by a one pass algorithm which sequences jobs one at a time from last to first. Suppose that jobs have been assigned to positions $k + 1$ through n. Let N_k be the set of jobs as yet unsequenced and L_k be the subset of N_k that can be assigned position k without violating the constraint. A job from L_k, say job j, is then chosen according to a certain rule and sequenced in position k. Then $N_{k-1} = N_k - \{j\}$, L_{k-1} is generated, and a job is sequenced in position $k - 1$, etc.

Letting $E_k = N_k \bigcap E$ and $p(N_k) = \sum_{i \in N_k} p_k$, then for problems (a) and (b)

$$L_k = \{N_k - E_k\} \cup \{j | j \in E_k;\ d_j \geqslant t(p(N_k))\}$$

while for problem (c)

$$L_k = \{j | j \in N_k;\ g_j(t(p(N_k))) \leqslant G_m\}$$

The rule for choosing the job for position k in problems (a) and (c) is choose j such that

$$p_j = \max_{i \in L_k} p_i$$

while for problem (b), j is chosen such that

$$d_j = \max_{i \in L_k} d_i.$$

Problem (d) may be solved by modifying the due-dates to reflect the fact that if $d_i \leqslant d_k$, $k \in E$, and job i is to be on time in a feasible sequence then i must be completed by $t(R(d_k) - p_k)$. Then Moore's algorithm can be applied, with an adjustment to assure that jobs in E will be on time. This is essentially the procedure developed by Sidney [15].

Nonsimultaneous Arrivals

In the preceding sections all jobs are assumed to be available for sequencing at time zero. We now consider problems in which job j is not available for processing until the beginning of period r_j, where $r_j \geqslant 1$. If, in this situation, it is possible to interrupt the processing of a job and resume it later without loss of progress toward completion of the job, we say that the system operates in a "prempt-resume" mode.

For single-machine problems with criteria maximum tardiness ($G = \max_{j \in N} T_j$) or total (mean) completion time ($G = \sum_{j \in N} C_j$) when prempt-resume applies; the static optimizing rules EDD and SPT can be generalized in a straightforward manner to produce optimal sequences when all jobs are not simultaneously available ([1] p. 82). The same generalizations apply when resource availability varies with time, using the following procedure:

1. At time zero if one or more jobs are available assign the resource to process the available job with the smallest (most urgent) priority. Otherwise leave the resource idle until the first job is available.

2. At each job arrival, compare the priority of the newly available job j with the priority of the job currently being processed. If the priority of job j is less, allow job j to preempt the job being processed; otherwise add job j to the list of available jobs.

3. At each job completion, examine the set of available jobs and assign the resource to process the one with the smallest priority.

In order to minimize maximum tardiness, the priority of a job is taken to be its due-date, and to minimize mean flowtime the priority is its remaining resource requirement.

3. MINIMIZING THE SUM OF WEIGHTED COMPLETION TIMES

One case for which the single-machine result does not generalize in a straightforward manner to the corresponding variable-resource problem is the case sequencing to minimize the sum of weighted completion times, where

$$G = \sum_{j \in N} w_j C_j$$

when all jobs are available at time zero.

Sequencing jobs in nondecreasing ordef of the ratio p_j/w_j, which will always minimize G in the single-machine problem, need not yield an optimal sequence when the resource availability varies with time. The following simple example demonstrates this fact.

EXAMPLE

j	1	2	3
p_j	7	3	6
w_j	5	2	4
p_j/w_j	1.4	1.5	1.5

$r(t) = 1 \quad 0 \leqslant t \leqslant 4$

$r(t) = 4 \quad 4 \leqslant t \leqslant 7 \quad M = 7$

Sequencing the jobs in nondecreasing order of p_j/w_j yields the order 1-2-3, for which the completion times are 4.75, 5.5 and 7. Therefore, $G = 62.75$. For the sequence 2-1-3 the completion times are 3, 5.5 and 7, with $G = 61.5$. (Under the discrete time framework $G = 65$ for 1-2-3 but $G = 64$ for 2-1-3.)

While the differences in G-values may seem almost insignificant it is possible to construct an example in which sequencing by increasing ratios p_j/w_j will yield an arbitrarily bad solution. Consider the data for a two-job problem in which

j	1	2
p_j	10^m	5×10^{2m}
w_j	1	10^m
p_j/w_j	10^m	5×10^m

$r(t) = 5 \times 10^{2m}, \quad 0 \leqslant t \leqslant 1$

$r(t) = 1 \qquad , \quad 1 < t \leqslant 1 + 10^m.$

Letting S represent the sequence 1-2 $(p_1/w_1 \leqslant p_2/w_2)$ and S', the sequence 2-1, for large m we have

$$\frac{G(S)}{G(S')} \cong \frac{1}{2}(10^m).$$

For the special case in which the processing times and weights are *agreeable* ($p_j \leqslant p_i$ whenever $w_j \geqslant w_i$) sequencing by nondecreasing ratios of p_j/w_j does produce an optimal solution (see Theorem 4). Otherwise the two examples given in this section reinforce the notion that the single-machine result cannot be extended to even the simplest versions of the variable-resource model. In one example the resource profile $r(t)$ is nondecreasing, while in the other example $r(t)$ is nonincreasing. In both cases there is only one change in $r(t)$. These situations would appear to be among the least drastic ways of relaxing the constant resource assumption; but, as we have demonstrated, the ratio rule still fails. At this point, we can conclude only that the minimization of $\Sigma w_j C_j$ involves more than a simple extension of the single-machine result. Obviously, any optimal ordering rule (if one exists) would have to involve information about the resource profile as well as information about processing requirements and weights. We conjecture that this problem is NP-complete.

4. COMMENTS

Although it is not possible to extend all single-machine results directly to the variable-resource case, a few observations can be made. A look at Figure 1 indicates that the graph of $R(t)$ transforms processing times (on the horizontal axis) into resource consumptions (on the vertical axis), and vice-versa. This transformation is at least order-preserving. In particular, the makespan for a set A of jobs is at least as large as the makespan for set B when the jobs in A have a total processing requirement that equals or exceeds the requirement of the jobs in B. This property is fundamental to the proof of many single-machine results as they carry over to variable-resource models. Moreover, the results for problems in which $p_j = p$ do not rely on the precise nature of the transformation, but depend only on the fact that all solutions share a common nondecreasing sequence of completion times.

In the single-machine case, $R(t)$ is linear, implying that the mapping of resource consumptions into processing times is proportionality-preserving as well as order-preserving. That is, ratios of intervals on the resource axis convert to identical ratios on the time axis. This property is not maintained in the variable-resource model, because the transformation distorts proportionality. In particular, we have in the single-machine problem that $p_i/p_j \leqslant w_i/w_j$ implies $\Delta C_i/\Delta C_j \leqslant w_i/w_j$, where ΔC_i and ΔC_j denote the magnitude changes in the completion times of adjacent jobs i and j which are interchanged in sequence. This implication does not hold in the variable-resource problem, so the pairwise interchange argument may fail.

These observations lead to the conclusion that single-machine results involving minimum weighted sum of completion times cannot be directly extended. An open question is therefore how to exploit the structure of this problem in the variable-resource case in order to find optimal solutions.

REFERENCES

[1] Baker, K.R., *Introduction to Sequencing and Scheduling*, Wiley (1974).

[2] Burns, R.N. and K.J. Noble, "Single Machine Sequencing with a Subset of Jobs Completed on Time," Working Paper, University of Waterloo, Canada (1975).

[3] Emmons, H., "One Machine Sequencing to Minimize Certain Functions of Job Tardiness," Operations Research, *17*, 701-715 (1969).

[4] Emmons, H., "One Machine Sequencing to Minimize Mean Flow Time with Minimum Number Tardy," Naval Research Logistics Quarterly, *22*, 585-592 (1975).

[5] Emmons, H., "A Note on a Scheduling Program with Dual Criteria," Naval Research Logistics Quarterly, *22*, 615-616 (1975).

[6] Gelders, L. and P. Kleindorfer, "Coordinating Aggregate and Detailed Scheduling in the One Machine Job Shop: Part I," Operations Research, *22*, 46-60 (1974).

[7] Gelders, L. and P. Kleindorfer, "Coordinating Aggregate and Detailed Scheduling in the One-Machine Job Shop: Part II," Operations Research, *23*, 312-324 (1975).

[8] Heck, H. and S. Roberts, "A Note on the Extension of a Result on Scheduling with a Secondary Criteria," Naval Research Logistics Quarterly, *19*, 403-405 (1972).

[9] Lawler, E.L., "Optimal Sequencing of a Single Machine Subject to Precedence Constraints," Management Science, *19*, 544-546 (1973).

[10] Lawler, E.L., "On Scheduling Problems with Deferral Costs," Management Science, *11*, 280-288 (1964).

[11] Lawler, E.L., "A Pseudopolynomial Algorithm for Sequencing Jobs to Minimize Total Tardiness," Annals of Discrete Mathematics, *1*, 331-342 (1977).

[12] Lawler, E.L. and J.M. Moore, "A Functional Equation and Its Application to Resource Allocation and Sequencing Problems," Management Science, *16*, 77-84 (1969).

[13] Moore, J.M. "An *n* Job, One Machine Sequencing Algorithm for Minimizing the Number of Late Jobs," Management Science, *15*, 102-109 (1968).

[14] Schrage, L.E. and K.R. Baker, "Dynamic Programming Solution of Sequencing Problems with Precedence Constraints," Operations Research, *26*, 444-449 (1978).

[15] Sidney, J.B., "An Extension of Moore's Due Date Algorithm," Symposium on the Theory of Scheduling and Its Application, (S.E. Elmaghraby, editor) Lecture Notes on Economics and Mathematical Systems 86, Springer-Verlag, Berlin, 393-398 (1973).

EVALUATION OF FORCE STRUCTURES UNDER UNCERTAINTY

Charles R. Johnson

Department of Economics and Institute for Physical Science and Technology
University of Maryland
College Park, Maryland

Edward P. Loane

EPL Analysis
Olney, Maryland

ABSTRACT

A model, for assessing the effectiveness of alternative force structures in an uncertain future conflict, is presented and exemplified. The methodology is appropriate to forces (e.g., the attack submarine force) where alternative unit types may be employed, albeit at differing effectiveness, in the same set of missions. Procurement trade-offs, and in particular the desirability of special purpose units in place of some (presumably more expensive) general purpose units, can be addressed by this model. Example calculations indicate an increase in the effectiveness of a force composed of general purpose units, relative to various mixed forces, with increase in the uncertainty regarding future conflicts.

INTRODUCTION

In planning the procurement of major weapons systems (submarines, aircraft, ships, etc.), an argument, based upon relative cost-effectiveness in certain uses, may be made for the development and purchase of some items which are less versatile and effective than the "best" available components of an overall force. Assuming all relative costs and effectivenesses known, such an argument is sound at least to the extent that the uses necessitated by a potential conflict are anticipated. However, under uncertainty about the nature of potential conflicts, a question, in general more subtle, is raised regarding the optimal composition of forces. In this case, a model is developed here to analyze the utility of "mixed" force structures, and examples are given to support the intuitive notion that the less specific are the presumptions about needs in a future conflict, the more valuable are the most versatile forces.

Our focus here is upon presenting a model able to capture the value, under uncertainty, of versatile forces and not upon the equally important problem of determination of cost and effectiveness parameters. The latter, as well as the mixture versus force level interaction, are touched upon tangentially in an example. The parameter estimation problem, in general, requires both large scale theoretical and empirical effort and has been addressed, in the submarine case, in Reference [1].

By *general purpose forces* we shall mean the most versatile, advanced or effective components which technology would currently allow in building a military force structure. *Special purpose forces*, on the other hand, might be competitive in effectiveness with general purpose, but only in some of the uses (which we shall call missions) which possible conflicts might require. Naturally, we presume that the general purpose are more expensive than the special purpose forces per item, and further that the special purpose forces are cost effective, in some missions. It is assumed also that all costs are accounted for, e.g., development, production, maintenance, operation, repair and logistical mobility, etc.

Examples of general versus special purpose forces include the following. In the case of submarine forces, the general purpose would be the newest fully equipped nuclear submarine while a special purpose alternative would be the conventional diesel submarine found in many European navies. The former is presumed at least as effective in all missions (much more so in some) while the latter is much less expensive and nearly as effective in some missions requiring only low mobility. In the case of aircraft, a long-range fighter-bomber might be considered general purpose and a plane designed primarily for ground attack would be special purpose.

The force planner must procure some mixture of forces, constrained, presumably, by a fixed budget. In general there may be several force types, ranging from the very general to the very special purpose, and we may think of the force structure as being a vector of inventories of each type purchased. We think of a conflict as simply a collection of mission opportunities, and the planner's problem is then to procure that force structure which permits the most effective deployment for a conflict. For a specified conflict, this poses a deterministic optimization problem which, if the conflict includes enough important mission opportunities in which the special purpose forces are cost effective, will surely suggest a mixed force structure including at least some special purpose units.

However, procurement of weapons systems must generally be decided upon long in advance of potential conflicts. For a variety of additional reasons, there will likely be considerable uncertainty as to the precise nature of an actual conflict. We consider this uncertainty to be characterized by a (known) distribution of potential conflicts, i.e., a distribution of mission opportunities. We note that there are other ways in which uncertainty might be treated. For example, if one's own force structure is known, a hostile adversary might be expected, to the extent that circumstances allow, to bias a conflict in a direction which would render one's own force least effective. This suggests a game theoretic approach. Although it is not pursued further here and although its information requirements might be great, this would naturally fit into the model context we outline below. It seems likely that such a treatment would value the versatility of general purpose forces more so than the one we pursue. Another alternative would be to treat the effectiveness of each unit type as unknown and characterize it by a probability distribution.

The planner's problem which we address is then to choose that affordable mixture of forces which, assuming optimal deployment in any conflict, yields the largest expected effectiveness in the uncertain conflict. It should be noted that, as stated, there is an implicit assumption that the planner is willing to take the risk that the solution mixture will produce unusually low effectiveness in some conflicts. (This is in contrast with the game theoretic approach mentioned above.) However, to the extent that the planner is risk-adverse rather than risk-neutral, other criteria may be substituted for "expected effectiveness" without conceptual difficulty and probably without operational difficulty in the development below. It should also be mentioned that a measure of the value of the versatility of general purpose forces under uncertainty lies in comparing the solution mixture of the above problem to the optimal mixture when the expected conflict is assumed known (i.e., the case of certainty). In general the "expected effectiveness" solution will differ from the "expected conflict" solution.

MODEL DESCRIPTION

We imagine n *force types* T_j, $j = 1, \ldots, n$ and m different *mission categories* U_i, $i = 1, \ldots, m$, in which a component of the force might be engaged. Each T_j is more or less effective in a given U_i which, to the extent that total effectiveness is linear in the deployment of force types to mission categories, suggests the definition of an m-by-n *unit effectiveness matrix*

$$E = (e_{ij}),$$

in which e_{ij} indicates the effectiveness of a unit of T_j employed in U_i for a unit of conflict (presuming opportunities available). We denote by a 1-by-n vector s, a particular *force composition* in which s_j is the number of T_j available. At the time of a conflict, s is fixed and, therefore, provides a constraint on the total effectiveness attainable. A particular conflict is characterized by the *total opportunity for effectiveness* which may be obtained from *each mission category.* These bounds are summarized in an m-by-1 vector b in which b_i is the maximum opportunity in U_i. This bound is expressed in effectiveness units rather than force units because the "opportunities" are opportunities to damage the opponent and the force types vary in their ability to do so in a given mission.

The m-by-n matrix $A = (a_{ij})$ summarizes the *allocation* (or deployment) of T_j to U_i, i.e., a_{ij} is the amount of force type T_j allocated to mission category U_i during a conflict. The a_{ij} are necessarily nonnegative but we do not assume them integral because of the possibility of switching units among missions.

The problem of waging a given conflict is then to deploy the given force so as to maximize total effectiveness within the constraint of the opportunities the conflict presents. In general (no linearity assumption), total effectiveness is some function

$$e = e(A)$$

of the allocation, and, furthermore,

$$e(A) = e_1(A) + \ldots + e_m(A),$$

where $e_i(A)$ is the effectiveness A yields through the ith mission category. This means that waging the known conflict b amounts to the optimization problem:

$$\begin{aligned} &\text{maximize} && e(A) \\ &\text{subject to} && \sum_{i=1}^{m} a_{ij} \leqslant s_j, \quad j = 1, \ldots, n \\ & && e_i(A) \leqslant b_i, \qquad i = 1, \ldots, m \\ & && a_{ij} \geqslant 0. \end{aligned}$$

In case total effectiveness is linear in A, we have the linear programming problem:

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} e_{ij} \\ &\text{subject to} && \sum_{i=1}^{m} a_{ij} \leqslant s_j, \quad j = 1, \ldots, n \\ & && \sum_{j=1}^{n} a_{ij} e_{ij} \leqslant b_i, \quad i = 1, \ldots, m \\ & && a_{ij} \geqslant 0. \end{aligned}$$

In either case we denote the maximum achieved by $M(s,b)$. Then, equicost force compositions s may be compared, for a given conflict, by comparing the $M(s,b)$. A good general reference for relevant concepts in the linear case is Reference [2].

Uncertainty as to the nature of the conflict is characterized by a probability distribution for b. For a given s, there is an $M(s,b)$ for each possible value of b. These may then be averaged according to the distribution of b to obtain the expected value:

$$M(s) = E_b(M(s,b)).$$

Comparisons among force compositions may then be made by comparing the $M(s)$, and the planner's problem is to

$$\text{maximize } M(s)$$

subject to his budget constraint governing the possible forces s which may be purchased. In general,

$$\max_s E_b(M(s,b)) \neq \max_s M(s,E_b(b)),$$

and in the case that effectiveness is linear in A,

$$\max_s E_b(M(s,b)) \leqslant \max_s (M(s,E_b(b)).$$

Thus, the maximum expected effectiveness problem has a different solution from the problem of maximum effectiveness is an expected conflict, so that uncertainty makes a difference in planning. We present examples which illustrate this, and in which the latter favors special purpose forces while the former favors general purpose, presumably because of their greater ability to defend against variation (uncertainty). The suggestion is that the more uncertainty there is, the greater the value of general purpose forces.

EXAMPLES

We conclude by giving two examples. The first is primarily to illustrate the evaluation model and some of the remarks made. The second includes a more thorough examination of the model and its assumptions in a detailed example intended to be suggestive of a realistic case which motivated this study.

EXAMPLE 1: Here we imagine three force types. Type T_1 is the general purpose, and T_2 and T_3 are different special purpose forces. There are also three mission categories. Type T_2 is cost effective relative to T_1 in mission U_1, while T_3 is cost effective relative to T_1 in U_3. Total effectiveness is assumed linear in allocations and the unit effectiveness matrix is

$$E = \begin{bmatrix} 1 & .7 & .1 \\ 1 & .1 & .1 \\ 1 & .1 & .7 \end{bmatrix}.$$

We consider seven equicost force compositions

$$\begin{aligned} s^1 &= (9,\ 0,\ 0) \\ s^2 &= (6,\ 3,\ 3) \\ s^3 &= (6,\ 6,\ 0) \\ s^4 &= (6,\ 0,\ 6) \\ s^5 &= (5,\ 4,\ 4) \\ s^6 &= (5,\ 8,\ 0) \\ s^7 &= (5,\ 0,\ 8). \end{aligned}$$

Thus, the two special purpose forces cost half as much as the general purpose over the range of procurement considered. (Actually, the outcome will not differ qualitatively if more alternatives based upon the 2-for-1 trade-off are considered.)

There are six possible conflicts

$$b^1 = \begin{pmatrix} 0 \\ 6 \\ 6 \end{pmatrix} \; b^2 = \begin{pmatrix} 6 \\ 6 \\ 0 \end{pmatrix} \; b^3 = \begin{pmatrix} 6 \\ 0 \\ 6 \end{pmatrix} \; b^4 = \begin{pmatrix} 12 \\ 0 \\ 0 \end{pmatrix} \; b^5 = \begin{pmatrix} 0 \\ 12 \\ 0 \end{pmatrix} \text{ and } b^6 = \begin{pmatrix} 0 \\ 0 \\ 12 \end{pmatrix}$$

with the first three presumed to have probability 2/9 each and the last three probability 1/9 each. Thus, the expected conflict is

$$\bar{b} = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}.$$

Straightforward calculations then yield

$$M(s^1) = 9$$
$$M(s^2) = M(s^3) = M(s^4) = 8.6, \text{ and}$$
$$M(s^5) = M(s^6) = M(s^7) = 8.47$$

so that

$$\max_{1 \leqslant i \leqslant 7} M(s^i) = 9$$

is achieved at s^1, the all general purpose force. On the other hand,

$$M(s^1, \bar{b}) = 9$$

while

$$M(s^2, \bar{b}) = 10.2, \; M(s^3, \bar{b}) = M(s^4, \bar{b}) \cong 10.1,$$
$$M(s^5, \bar{b}) = 10.6, \text{ and } M(s^6, \bar{b}) = M(s^7, \bar{b}) \cong 9.3.$$

Thus, a mixed force s^5 is optimal for the expected conflict. The conclusion, in this case, is that general purpose forces are overall more cost efficient under uncertainty. It should be noted that in calculating the $M(s^i)$, each other force had higher effectiveness than s^1 for some conflicts (but not overall) and all were better than s^1 in the expected conflict. Thus, it is only the value of versatility under uncertainty which makes s^1 preferred.

EXAMPLE 2: This example is taken from the problem of submarine procurement and again illustrates the effect of uncertainty on the attractiveness of special purpose forces.

For simplicity, we consider only two types of forces, general purpose and special purpose units. In this setting, the distinction between new procurement general purpose or special purpose forces might well be that between nuclear or diesel-electric propulsion. Equipment and weapons could be identical, but the lower underwater mobility inherent in diesel-electric propulsion would limit effective employment of such forces to particular ASW missions. In the actual planning process, the existing force structure must also be considered since in a future conflict, presently existing units might be restricted to low vulnerability missions (presumably being less capable than new procurement general purpose units) and thus constitute additional categories of special purpose forces.

The present example considers four missions and measures unit effectiveness in each mission by a *kill rate* defined by:

$$e_{ij} = \frac{\left[\begin{array}{l}\text{Kills (of enemy submarines) per unit}\\ \text{time by one on-station U.S. submarine}\\ \text{of type } T_j \text{ engaged in mission } U_i\end{array}\right]}{[\text{Number of surviving enemy submarines}]}.$$

The above quantity is well defined for important submarine missions, being independent of enemy force size and the number of U.S. submarines committed to U_i over a substantial range of values. For instance, considering a fixed barrier mission, the rate of enemy transits through the barrier and thus the rate of opportunities for kill would be proportional to the number of surviving units. Also, U.S. submarine probabilities of detection and kill given an opportunity (here target passage through the barrier area assigned to the submarine) are, at least initially, inversely proportional to the width of the barrier area assigned. In this circumstance, e_{ij} is well defined. Of course, nonlinear effects are present and become significant as the number of U.S. units is increased. One could argue that, as returns diminish, no additional submarines should be assigned to the fixed barrier; this then determines the mission opportunities, b_j. With units of differing capabilities, b_j is properly stated in terms of effectiveness obtained, not in some fixed maximum number of units employed, since the onset of diminishing returns would occur at different force levels for different unit effectivenesses. Finally, variations in b_j (for the fixed barrier mission) might arise from uncertainties in enemy basing, at-sea replenishment of submarines, desirable barrier locations being untenable due to enemy ASW, etc.

Similar arguments apply for the direct support mission (submarines employed in the defense of surface formations) and similar conclusions are obtained in the area search mission.

It should be noted that kill rates add, and that the summation

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} e_{ij},$$

being an overall rate at which enemy submarines are being killed, is a sensible measure of effectiveness for the entire U.S. submarine force. It is even plausible that the differing submarine types would be assigned to missions so as to (approximately) maximize this sum. Finally, to the extent that variations in b_j reflect week-to-week changes within a single conflict (i.e., one week large numbers of forces are required for direct support, the next week these same units are used in a barrier) rather than uncertainty as to some long-term mix of missions that will be required in an unspecified conflict, then the expected value

$$E_b(M(s,b))$$

can be interpreted as a time-average of force kill rate and again this is a preeminently sensible measure.

It is the authors' belief that the use of kill rates as measures of unit effectiveness and the linear formulation of force effectiveness, while necessarily involving some approximation, does capture the important aspects of evaluating alternative submarine force structures. Of course, in realistic applications, the evaluation of effectiveness for alternative forces is a substantial effort. Reference [1] documents a major study effort which arrives at such estimates, although not expressed as kill rates. Evaluation of force effectiveness is not addressed here. Quantitative inputs to this second example, shown in the following tabulation, are completely hypothetical; and, while of reasonable relative magnitudes, are chosen to illustrate the theses of this paper.

TABLE 1.

Unit Effectiveness, e_{ij} (Kill rates)			
	General Purpose Submarines	Special Purpose Submarines	Expected Total Opportunity for Effectiveness $E_b(b_i)$
Mission 1	1.0	.95	16
Mission 2	1.50	.50	16
Mission 3	.75	.375	12
Mission 4	.40	.20	Unlimited

TABLE 2.

Alternative Force Compositions, (s_1,s_2) (Numbers of Units on-station)	
General Purpose Submarines	Special Purpose Submarines
35	0
25	10
20	17
15	24

Unit effectivenesses and force compositions are stated in terms of on-station submarines; actual numbers of operational units would be higher than, and not necessarily in proportion to, the numbers shown. The alternative forces shown might well be equal cost options if there were some fixed cost associated with deploying any special purpose submarines. The fourth mission is not limited in the number of forces which can be employed or the total effectiveness which can be obtained. This might be thought of as undirected open-ocean search, which could always be undertaken by any submarine not otherwise assigned.

The distributions of b, reflecting uncertainty, are represented by lists of 60 sample vectors—each considered equally likely. The lists are not repeated here. Sample vectors were generated by Monte-Carlo methods, assuming each b_i is an independent truncated* Gaussian random variable with the above stated mean and relative standard deviations of 35% and 60% in the two cases considered. Effectiveness, for the alternative force compositions is shown in Table 3, following.

The maximal effectiveness for each level of uncertainty is enclosed in dashes. Not surprisingly, the example values show a change in preference, from a mixed force to an all general purpose force, as variability in mission opportunities increases. What is surprising is that the changes, and differences are so small overall. This can be explained qualitatively, and is a reflection of a real concern in procurement decisions.

*Both high and low values were discarded so as to preserve the mean value and assure that $b_i \geqslant 0$.

TABLE 3.

Force Compositions, s	Force Effectiveness, $M(s)$		
(s_1, s_2)	No Uncertainty (mean value b used)	Relative Standard Deviation of each $b_i = 35\%$	Relative Standard Deviation of each $b_i = 60\%$
(35, 0)	38.2	37.4	[36.5]
(25, 10)	37.9	36.9	35.8
(20, 17)	[39.1]	[37.6]	36.2
(15, 24)	37.9	37.1	36.0

In the present example, the attractiveness of special purpose units rests on the availability of opportunities in mission 1; i.e., if $b_1 \geqslant 11.4$ then forces including some special purpose units are preferred to an all general purpose force. But mission 1 is a substantial (36%) of the projected employment of submarines; if this were taken away, then the force is over-built and any alternative composition is able to exploit the remaining attractive opportunities. That is, if $b_1 \rightarrow 0$ then all force compositions entertained give about the same effectiveness; and as noted above, if $b_1 \geqslant 11.4$, compositions involving special purpose units are preferred. In this circumstance, i.e., with the numeric inputs to this example calculation, one cannot expect to see dramatic changes in preferences among force compositions, with explicit consideration of uncertainty.

As a final point, we note the suboptimality of separating questions of force composition from questions of force levels. Although this raises an issue worthy of further study, we only mention the issue here by extending the previous example. Using exactly the same unit effectiveness and mission opportunity values stated previously, but considering alternative force compositions which involve an additional 5 general purpose submarines, one obtains the following results:

TABLE 4.

Force Compositions, s	Force Effectiveness, $M(s)$		
(s_1, s_2)	No Uncertainty (mean value b used)	Relative Standard Deviation of each $b_i = 35\%$	Relative Standard Deviation of each $b_i = 60\%$
(40, 0)	42.0	40.7	39.6
(30, 10)	41.6	40.3	39.0
(25, 17)	[42.8]	[41.0]	39.6
(20, 24)	41.7	40.7	[39.7]

In this case, the uncertainty considered does not lead to a preference for an all general purpose force, although again the effects are very small. The tendency here is intuitively satisfying, i.e., special purpose units become more attractive as overall force levels are increased, relative to a fixed job to be done. Notice also that increased uncertainty decreases the incremental effectiveness of the additional five general purpose units, in every case.

REFERENCES

[1] Chief of Naval Operations, Future Submarine Employment Study (U), (29 December 1972)-SECRET.

[2] Dantzig, G.B., *Linear Programming and Extensions,* Princeton University Press (1963).

A NOTE ON THE OPTIMAL REPLACEMENT TIME OF DAMAGED DEVICES

Dror Zuckerman

The Hebrew University of Jerusalem
Israel

ABSTRACT

Abdel Hameed and Shimi [1] in a recent paper considered a shock model with additive damage. This note generalizes the work of Abdel Hameed and Shimi by showing that the *a*-priori restriction to replacement at a shock time made in [1] is unnecessary.

1. INTRODUCTION

A recent paper by Abdel Hameed and Shimi [1] was concerned with determining the optimal replacement time for a breakdown model under the following assumptions: A device is subject to a sequence of shocks occurring randomly according to a Poisson process with parameter λ. Each shock causes a random amount of damage and these damages accumulate additively. The successive shock magnitudes $Y_1, Y_2, \ldots$, are positive, independent, identically distributed random variables having a known distribution function $F(\cdot)$. A breakdown can occur only at the occurrence of a shock. Let δ denote the failure time of the device. For $t < \delta$ let $X(t)$ be the accumulated damage over the time duration $[0,t]$. The device fails when the accumulated damage $X(t)$ first exceeds Z. That is,

$$\delta = \inf\{t \geqslant 0;\ X(t) \geqslant Z\}, \tag{1}$$

where Z is a random variable, independent of the accumulated damage process X, having a known distribution function $G(\cdot)$ called the killing distribution. More explicitly, if $X(t) = x$ and a shock of magnitude y occurs, at time t, then the device fails with probability

$$\frac{G(x+y) - G(x)}{1 - G(x)}. \tag{2}$$

Upon failure the device is immediately replaced by a new identical one with a cost of c. When the device is replaced before failure, a smaller replacement cost is incurred. That cost depends on the accumulated damage at the time of replacement and is denoted by $c(x)$. That is to say $c(x)$ is the cost of replacement before failure when the accumulated damage equals x. It is assumed that $c(0) = 0$ and $c(x)$ is bounded above by c. Thus there is an incentive to attempt to replace the device before failure. The condition $c(0) = 0$ has to be interpreted as a policy of no replacement if there is no damage.

In their paper Abdel Hameed and Shimi [1] derived an optimal replacement policy that minimizes the expected cost per unit time under the restriction that the device can be replaced only at shock point of time.

In the present article we consider a similar breakdown model without the above restriction made in [1]. We allow a controller to institute a *replacement* at any stopping time before failure time. He must replace upon device failure. Throughout, we restrict attention to replacement policies for which cost of replacement is solely a function of the accumulated damage. In some shock models, replacement at a scheduled time offers potential benefits relative to replacement at a random time. However, the problem of scheduled replacement in failure models with additive damage is an open problem and it is beyond the scope of the present study.

Let T be the replacement time. At time T the device is replaced by a new one having statistical properties identical with the original, and the replacement cycles are repeated indefinitely. The collection of all permissible replacement policies described above will be denoted by M. Our objective is to prove that an optimal policy replaces the system at shock point of time. Thus the restriction about the class of permissible replacement policies made in [1] can be omitted.

The following will be standard notation used throughout the paper: $E[Y;A]$, where Y is a random variable and A is an event, refers to the expectation $E[I_A Y] = E[Y|I_A = 1]P(A)$, where I_A is the set characteristic function of A.

2. THE OPTIMAL POLICY

By applying a standard renewal argument, the long run average cost per unit time when a replacement policy T is employed can be expressed as follows

$$\psi_T = \frac{E[c(X(T));\ T < \delta] + E[c;\ T = \delta]}{E[T]}. \tag{3}$$

Let $\psi^* = \inf_{T \in M} \psi_T$.

Clearly

$$\psi^* \leqslant \frac{E[c(X(T));\ T < \delta] + E[c;\ T = \delta]}{E[T]},$$

for every $T \in M$, and the optimal replacement policy that minimizes ψ_T over the set M is the one that maximizes

$$\theta_T = \psi^* E[T] + E[c - c(X(T));\ T < \delta]. \tag{4}$$

By applying Dynkin's formula (see Theorem 5.1 and its Corollary in Dynkin [2]) equation (4) reduces to

$$\theta_T = E\left[\int_0^T J(X(s))\,ds\right] + c, \tag{5}$$

where

$$J(x) = \psi^* - \lambda c\left\{1 - \int \frac{\bar{G}(x+y)}{\bar{G}(x)}\,dF(y)\right\} + \lambda\left\{c(x) - \int c(x+y)\,\frac{\bar{G}(x+y)}{\bar{G}(x)}\,dF(y)\right\}. \tag{6}$$

The proof of the above result follows a procedure similar to that used by the author in (Section 2 of [3]), and therefore is omitted.

In what follows we shall denote by S the state space of the stochastic process $\{X(t);\ t < \delta\}$.

Let

$$S_1 = \{x \in S;\ J(x) > 0\}, \tag{7}$$

and

$$S_2 = \{x \in S;\ J(x) \leqslant 0\}. \tag{8}$$

Let $t_1, t_2, t_3, \ldots,$ be the shock points of time and define

$$W = \{t_i;\ i \geqslant 1\}.$$

Let L be the subclass of replacement policies in which a decision can be taken only over the set W.

We proceed with the following result:

THEOREM 1: For every replacement policy $T_1 \notin L$, there exists a replacement policy $T_2 \in L$ such that $\theta_{T_2} \geqslant \theta_{T_1}$.

PROOF: Let T_1 be a replacement policy such that $T_1 \notin L$.

Let $T(S_2)$ be the hitting time of the set S_2. That is

$$T(S_2) = \inf\{t \geqslant 0;\ X(t) \in S_2\}. \tag{9}$$

(It is understood that when the set in braces is empty, then $T(S_2) = \infty$.)

Let

$$\tilde{T} = \inf\{t \geqslant T_1;\ t \in W\} \tag{10}$$

and define

$$T_2 = \min\{\tilde{T},\ T(S_2)\}. \tag{11}$$

Clearly $T_2 \in L$. Next we show that $\theta_{T_2} \geqslant \theta_{T_1}$.

Using (5) we obtain

$$\theta_{T_2} - \theta_{T_1} = E\left[\int_0^{T_2} J(X(s))\,ds\right] - E\left[\int_0^{T_1} J(X(s))\,ds\right] = E\left[\int_{T_1}^{\tilde{T}} J(X(s))\,ds;\ T_2 = \tilde{T}\right]$$

$$- E\left[\int_{T(S_2)}^{T_1} J(X(s))\,ds;\ \tilde{T} > T(S_2)\right]. \tag{12}$$

Note that

I. $\{T_2 = \tilde{T}\}$ implies that $\{T(S_2) \geqslant \tilde{T}\}$ and therefore $E\left[\int_{T_1}^{\tilde{T}} J(X(s))\,ds;\ T_2 = \tilde{T}\right] \geqslant 0$

II. $J(X(s))$ for $T(S_2) \leqslant s < T_1$ is *non-positive* on the set $\{\tilde{T} > T(S_2)\}$. Therefore $E\left[\int_{T(S_2)}^{T_1} J(X(s))\,ds;\ \tilde{T} > T(s_2)\right] \leqslant 0.$

Therefore, (using (12)) we obtain

$$\theta_{T_2} - \theta_{T_1} \geqslant 0$$

as desired.

Recalling that an optimal replacement policy T^* is the one that maximizes θ_T and using Theorem 1. it follows that $T^* \in L$. Hence, the optimal policy derived by [1] is the optimal one among all possible replacement policies for which cost of replacement is solely a function of the accumulated damage.

Finally it should be pointed out that if the benefits of scheduled replacement were considered, the conclusion reached, that an optimal policy replaces the device at a shock point of time, would no longer generally hold.

REFERENCES

[1] Abdel Hameed, M. and I.N. Shimi, "Optimal Replacement of Damaged Devices," Journal of Applied Probability *15,* 153-161, (1978).
[2] Dynkin, E.G., *Markov Processes I,* Academic Press, New York, (1965).
[3] Zuckerman, D. "Replacement Models under Additive Damage," Naval Research Logistics Quarterly, *24,* 549-558, (1977).

A NOTE ON THE SENSITIVITY OF NAVY FIRST TERM REENLISTMENT TO BONUSES, UNEMPLOYMENT AND RELATIVE WAGES*

Les Cohen

Government Services Division
Kenneth Leventhal & Company
Washington, D.C.

Diane Erickson Reedy

Mathtech, Inc.
Rosslyn, Virginia

ABSTRACT

Multiple regression analysis of first term reenlistment rates over the period 1968-1977 confirms previous findings that reenlistment is highly sensitive to unemployment at the time of reenlistment and shortly after enlistment, almost four years earlier. Bonuses, particularly lump sum bonuses, were also shown to be a significant determinant of reenlistment.

This note reports the results of cross-sectional multiple regression analysis of first term Navy reenlistment. Equations which were estimated represent the completion of research conducted by Cohen and Reedy [1] which analyzed the sensitivity of first term reenlistment to fluctuations in economic conditions at the time of reenlistment and about the time of enlistment, considering the effect of the latter on reenlistment behavior four years later. The principal finding of that study was that unemployment rates, both at the time of reenlistment and about the time of enlistment four years earlier, were powerful predictors of reenlistment rates. By comparison, measures of private sector versus military wages entered in the same equations were generally found to be insignificant or, at best, relatively unimportant. That study did not, however, take into account the influence of reenlistment bonuses which this follow-up note addresses.

This note describes the results of regression equations, replicating those which were the basis of the original Cohen-Reedy paper, which include reenlistment bonus variables to consider their influence upon Navy reenlistment over the ten year period, 1968-1977.

Reenlistment rates were compiled from *Navy Military Personnel Statistics* ("The Green Book"), quarterly by rating, separately for E-4's and E-5's. To help minimize spurious fluctuations in the data, reenlistment rates were calculated only for those quarters which had an average of at least 10 eligibles per month. In addition, due to definitional and mensurational inconsistencies, ratings which include nuclear power and diver NEC's were eliminated and other

*This research was supported by the Office of the Chief of Naval Operations, Systems Analysis Division, under a contract with Information Spectrum, Inc., Arlington, Virginia.

ratings which include 6 year obligors (6YO's) were analyzed separately. The resultant data base consisted of 3110 observations for 4YO ratings, and 787 observations for 6YO ratings. Each observation referred to a specific quarter, rating and pay grade, either E-4 or E-5.

Four multiple linear regression equations were estimated: one for 4YO ratings (including E-4's and E-5's); one for 6YO ratings (including E-4's and E-5's); one for 4YO E-4's; and one for 4YO E-5's. No attempt was made to estimate separate equations for each major occupational category as was done in the previous study. Given observed variations in earlier equations, collective treatment of ratings has probably resulted in depressed R^2 statistics.

The dependent variable, RATE3, is the percentage deviation of the current quarter reenlistment rate from the mean reenlistment rate for that rating and pay rate over the 10 years under study, 1968-1977.

$$\text{RATE3} = \frac{\text{(Quarterly Reenlistment Rate - Mean (10 Year) Reenlistment Rate)}}{\text{Mean (10 Year) Reenlistment Rate}}$$

This specification of the dependent variable was adopted to contend with wide variations in the level of reenlistment rates from rating to rating. RATE3 describes *relative* changes in reenlistment rates.

Independent variables included in the equations are listed and defined in Table 1.

TABLE 1 — *Independent Variables*

AUR	current national unemployment rate
ARAUR	average rate of change in unemployment (AUR) over the past 6 quarters preceeding the reenlistment decision
AUR13	unemployment (AUR) 13 quarters prior to the reenlistment decision (NOTE: Virtually uncorrelated with AUR.)
RW	the ratio of military basic pay to private sector earnings
AWARD	bonus award multiple
LS	dummy variable indicating lump sum payment of bonuses (LS = 1 for 1968 - 1974; LS = 0 for 1975 - 1977)
ELIG	number of individuals eligible for reenlistment
PAYRATE	dummy variable indicating rate (PAYRATE = 1 for E - 5's; PAYRATE = 0 for E - 4's)
DRAFT	number of persons drafted (all services) 18 quarters prior to reenlistment decision
WAR	dummy variable for Viet Nam War (WAR = 1 for 1968 - 1972; WAR = 0 for 1973 - 1977)
QTR3	third quarter seasonal dummy (QTR3 = 1 for 3rd calendar quarter only)
TIME	time variable (TIME = Year - 67)

In the context of cross-sectional analysis, estimated coefficients do not pertain to the impact of a given variable over time for a specific rating, but represent the typical impact of that variable over the entire 10 years across all ratings which were included in the study.

Results of the estimation procedures are summarized in Table 2.

TABLE 2 — *Reenlistment Equations: Coefficients, (t-statistics), and Means*

EQUATION	4YO		6YO		4YO/E-4		4YO/E-5	
INDEPENDENT VARIABLE	Coef. (*t*)	Mean	Coef. (*t*)	Mean	Coef. (*t*)	Mean	Coef. (*t*)	Mean
AUR	14.52	.06	18.46	.06	15.77	.06	12.49	.06
	(4.96)		(5.07)		(3.56)		(3.41)	
ARAUR	-.84	.02	-1.56	.02	-.70	.02	-1.38	.02
	(1.61)		(2.34)		(.88)		(2.12)	
AUR13	29.38	.04	25.13	.05	24.21	.04	36.87	.04
	(12.44)		(8.00)		(6.75)		(12.50)	
RW	-.63	.77	.48	.77	.94	.72	-.68	.82
	(1.12)		(.64)		(1.03)		(1.03)	
AWARD	.03	1.15	.0004	2.31	.03	1.15	.02	1.15
	(4.06)		(.05)		(3.47)		(2.43)	
LS	.45	.76	.57	.67	.31	.76	.50	.76
	(6.32)		(5.81)		(2.90)		(5.62)	
ELIG	- .08E-2	103.76	- .03E-2	165.82	- .07E-2	123.71	- .001	83.74
	(8.74)		(3.83)		(5.69)		(6.92)	
PAYRATE	.03	.50	- .05	.50				
	(.45)		(.66)					
DRAFT	- .04E-4	5.18E+4	- .04E-4	4.77E+4	- .04E-4	5.18E+4	- .04E-4	5.17E+4
	(6.69)		(4.82)		(5.03)		(5.51)	
WAR	.26	.57	.36	.50	.21	.57	.37	.57
	(4.10)		(4.46)		(2.12)		(4.61)	
QTR3	- .09	.26	- .08	.26	- .08	.26	- .07	.26
	(3.60)		(2.46)		(2.17)		(2.32)	
TIME	.08	5.05	.08	5.63	.08	5.05	.06	5.06
	(4.49)		(3.86)		(2.94)		(2.85)	
CONSTANT	- 2.26		- 3.03		- 3.10		- 2.39	
	(5.27)		(5.85)		(4.47)		(4.23)	
R^2	.34		.48		.40		.35	
OBSERVATIONS	3110		787		1556		1554	

The three unemployment variables, AUR, ARAUR and AUR13, were specified precisely as in the earlier Cohen-Reedy study. Consistent with those results, the significance of the unemployment rate variables and the magnitude of their apparent effect upon reenlistment are striking. Taken literally, coefficients in the 4YO equation, for example, show a one point increase in AUR13 (+.01) indicating a 29 point (+.29) increase in RATE3. While it is realized that these coefficients may overstate the real influence of unemployment, their equations, like those which they are replicating, do indicate that reenlistment decisions may in fact be sensitive to perceived costs of employment search and to the security of private sector employment.

The first compensation variable, RW, representing the ratio of military to private sector wages, was calculated separately for E-4's and E-5's using basic pay for E-5's and E-6's respectively as proxies for next-term earnings. RW was not a significant variable in any of the four equations.

The other two compensation variables, AWARD and LS, relate to bonuses. AWARD is the multiple for a particular rating in a given quarter, ranging from 0 to 6. This multiple is the factor which the Navy applies against an individual's monthly pay to compute the dollar amount of his bonus payment. AWARD was significant in all three 4YO equations. LS is a dummy variable which assumes a value of 1 through calendar 1974 during the period when lump sum awards were paid to approximately 50% of those individuals who reenlisted. Beginning January 1, 1975, a new policy was initiated which reduced the percentage of lump sum bonus payments to approximately 10% of those reenlisting. The coefficient of LS indicates that when bonuses were paid in lump sums, the percentage difference between actual reenlistment rates and mean (10 year) reenlistment rates was higher by .45 than when bonuses were paid in installments.

The variable ELIG was included in the equations simply to capture the observed relationship between low numbers of eligibles and high reenlistment rates.

PAYRATE is a dummy variable which distinguishes between E-4's and E-5's (PAYRATE = 1). TIME was included to capture the influence of factors which have changed steadily over time such as the quality of life improvements effected by the Navy over the past several years.

These equations support the authors' earlier findings, notably that unemployment rates at the time of the reenlistment decision and shortly after enlistment are important determinants of reenlistment rates. Relative wages continue to appear unimportant. It appears, however, that reenlistment bonuses have had a significant positive effect on reenlistment, particularly when those bonuses have been awarded in lump sum payments.

Although by no means conclusive, the equations summarized in Table 2 suggest the following management initiatives:

— Experimentation is warranted in the use of lump sum bonuses to mitigate the effects of low unemployment rates on reenlistment.

— Opportunities to reenlist might be timed to coincide with low points (periods of high unemployment) in the business cycle.

— AUR13 and predicted AUR should be used to augment current information used for projecting reenlistment rates.

— Based on the continued performance of the AUR13 variable, serious consideration must be given to implementing new programs designed to effect enlistee career decision making very early during the first term of service.

REFERENCES

[1] Cohen, L. and D. Reedy, "The Sensitivity of Navy First Term Reenlistment Rates to Changes in Unemployment and Relative Wages," Naval Research Logistics Quarterly, *26*, 695-709 (1979).

U.S. GOVERNMENT PRINTING OFFICE: 1980—311-493/3

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

SEPTEMBER
VOL. 27, NO. 3
NAVSO P-1278

CONTENTS

ARTICLES		Pa
On the Reliability, Availability and Bayes Confidence Intervals for Multicomponent Systems	W. E. THOMPSON R. D. HAYNES	3
Optimal Replacement of Parts Having Observable Correlated Stages of Deterioration	L. SHAW S. G. TYAN C-L. HSU	3
Statistical Analysis of a Conventional Fuze Timer	E. A. COHEN, JR.	3
The Asymptotic Sufficiency of Sparse Order Statistics in Tests of Fit with Nuisance Parameters	L. WEISS	3
On a Class of Nash-Solvable Bimatrix Games and Some Related Nash Subsets	K. ISAACSON C. B. MILLHAM	4
Optimality Conditions for Convex Semi-Infinite Programming Problems	A. BEN-TAL L. KERZNER S. ZLOBEC	4
Solving Incremental Quantity Discounted Transportation Problems by Vertex Ranking	P. G. MCKEOWN	
Auxiliary Procedures for Solving Long Transportation Problems	J. INTRATOR M. BERREBI	
On the Generation of Deep Disjunctive Cutting Planes	H. D. SHERALI C. M. SHETTY	
The Role of Internal Storage Capacity in Fixed Cycle Production Systems	B. LEV D. I. TOOF	
Scheduling Coupled Tasks	R. D. SHAPIRO	
Sequencing Independent Jobs With a Single Resource	K. R. BAKER H. L. W. NUTTLE	
Evaluation of Force Structures Under Uncertainty	C. R. JOHNSON E. P. LOANE	
A Note on the Optimal Replacement Time of Damaged Devices	D. ZUCKERMAN	
A Note on the Sensitivity of Navy First Term Reenlistment to Bonuses, Unemployment and Relative Wages	L. COHEN D. E. REEDY	

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217